Pareto Optimality and Strategy-Proofness in Group Argument Evaluation

Edmond Awad[1,2], Martin Caminada[3], Gabriella Pigozzi[4], Mikołaj Podlaszewski[5], and Iyad Rahwan[1,6,2,†]

[1]*The Media Lab, Massachusetts Institute of Technology, USA*
[2]*Masdar Institute, UAE*
[3]*School of Computer Science & Informatics, Cardiff University, UK*
[4]*Université Paris-Dauphine, PSL Research University, CNRS, LAMSADE, 75016 Paris, France*
[5]*University of Luxembourg, Luxembourg*
[6]*Institute for Data, Systems, & Society, Massachusetts Institute of Technology, USA*
[†]*Correspondence should be addressed to* irahwan@mit.edu

Abstract

An inconsistent knowledge base can be abstracted as a set of arguments and a defeat relation among them. There can be more than one consistent way to evaluate such an argumentation graph. Collective argument evaluation is the problem of aggregating the opinions of multiple agents on how a given set of arguments should be evaluated. It is crucial not only to ensure that the outcome is logically consistent, but also satisfies measures of social optimality and immunity to strategic manipulation. This is because agents have their individual preferences about what the outcome ought to be. In the current paper, we analyze three previously introduced argument-based aggregation operators with respect to Pareto optimality and strategy-proofness under different general classes of agent preferences. We highlight fundamental trade-offs between strategic manipulability and social optimality on one hand, and classical logical criteria on the other. Our results motivate further investigation into the relationship between social choice and argumentation theory. The results are also relevant for choosing an appropriate aggregation operator given the criteria that are considered more important, as well as the nature of agents' preferences.

1 Introduction

Argumentation has recently become one of the main approaches for non-monotonic reasoning and multi-agent interaction in artificial intelligence and computer science [6, 9, 37]. The most prominent approach in argumentation models is probably the abstract argumentation framework (AAF) by Dung [24], in which the contents of the arguments are abstracted from and the framework can be represented as a directed graph in which nodes represent arguments, and arcs between these nodes represent binary *defeat* relations over them. An important question is which arguments to accept. In his seminal paper, Dung has defined extension-based semantics which correspond to different criteria of acceptability of arguments. Another equivalent labeling-based semantics is proposed by Caminada [14]. Using this approach, an argument is labeled `in` (i.e. accepted), `out` (i.e. rejected), or `undec` (i.e. undecided). One of the essential properties, that is common, is the condition of *completeness*. Every complete (*i.e.* legal) labeling represents a consistent self-defending point of view. Since there can be different reasonable positions regarding the evaluation of an argumentation graph, choosing one legal labeling above another is not a trivial task. Therefore, in a multi-agent setting, different agents can subscribe to different positions. Hence, a group of agents with an argumentation graph would need to find a collective labeling that best reflects the opinion of the group. Despite the apparent simplicity of the problem, the aggregation of individual evaluations can result into an inconsistent group outcome. Recently, the problem of aggregating valid labelings has been the topic of some studies [38, 4, 16, 12, 10]. In the work by Caminada and Pigozzi [16], they proposed three possible operators for aggregating labelings, namely the skeptical operator, the credulous operator, and the super credulous operator. These operators guarantee not only a well-formed outcome but also a compatible one, that is, it does not go against the judgment of any individual. Recently, dialectical proof procedures have been stated by Caminada and Booth [15] for these three operators.

Although the outcomes of these three aggregation operators are compatible with every individual's labeling, this does not mean that they are the most desirable given individuals' preferences. It is possible that other compatible labelings are more desirable. Moreover, it is possible that some agents submit an insincere opinion in order to get more desirable outcomes. Given that, it is interesting to study the following two questions:

1. Are the social outcomes of the three aggregation operators Pareto optimal if preferences between different outcomes are also taken into consideration?

2. How robust are these operators against strategic manipulation? And what are the effects of strategic manipulation from the perspective of social welfare?

The first question studies the Pareto optimality of the outcomes of these operators. A Pareto optimal outcome (given individuals preferences) cannot be replaced with another outcome that is more preferred by all individuals and is strictly more preferred by at least one individual. Pareto optimality is a fundamental concept in any social choice setting and a clearly desirable property for any aggregation operator. The second question studies the strategy-proofness of the operators. Strategy-proofness is fundamental in any realistic multi-agent setting. A strategy-proof operator is

one that produces outcomes where individuals have no incentive to misrepresent their votes (i.e. to lie). Unfortunately, as we will see later, most strategy-proofness results for the three operators are negative. However, we show later that lies do not always have bad effects on other agents.

One can realize that individuals' preferences (over all the labelings) play a vital role in answering the previous two questions. However, aggregation operators usually do not give the chance for individuals to disclose these preferences. The labeling an agent submits is the only information available about agent's preferences. It seems a natural choice to assume that the submitted labeling is the most preferred one according to agents' individual preference. Moreover we assume that the rest of agent's preferences can be modeled using distance from the most preferred one. For example, if the top preferred outcome for agent i is the outcome O_1 (*i.e.* $\forall O_j$, $O_1 \succeq_i O_j$), then $O_2 \succ_i O_3$ iff $dist(O_1, O_2) < dist(O_1, O_3)$ where $dist(O_1, O_2)$ is the distance between the two outcomes O_1 and O_2. In this work, we investigate different classes of preferences based on different distance measures, and use them to analyze the three aggregation operators proposed by Caminada and Pigozzi [16] with respect to the aforementioned two questions.

This paper makes three distinct contributions. First, it introduces the first thorough study of Pareto optimality and strategy-proofness for aggregation operators in the context of argumentation. In doing so, the paper highlights that considering argumentation in multi-agent conflict resolution calls for criteria other than logical consistency such as social optimality and strategic manipulation. Second, the paper introduces different families of agents' preferences. For example, we define a new class of preferences which consider the label `undec` as a middle label between `in` and `out`. The variety of the different families of preferences are meant to broaden the scope of analysis of preferences, and test the robustness of the studied operators with respect to the considered questions. The third contribution of this paper is establishing relations between the different classes of preferences, and providing a full comparison for three previously introduced labeling aggregation operators with respect to the proposed classes of preferences. Moreover, cases where agents do not share the same classes of preference are also considered.

Our results bridge a gap in our understanding of the social optimality and strategic manipulation of labeling aggregation operators. As for the Pareto optimality, we show the persistence of the superiority of the skeptical operator. However, there are situations where the credulous and super credulous operators are as good as the skeptical operator. This has an implication on the choice of the appropriate aggregation operator given the criteria that is considered more important, as well as, the nature of agents preferences. As for the strategy-proofness, we establish the fragility of the three operators against strategic manipulation. This negative result is consistent even for a wide range of individual agent preference criteria (except for one case). This highlights a major limitation of these otherwise attractive approaches to collective argument evaluation. Despite the negative results, our results show that lies with the skeptical operator are always benevolent i.e. every strategic lie by an agent does not hurt others, but rather improves their welfare. Furthermore, we show that this effect is surprisingly consistent for a wide range of individual agent preference criteria. This shows an important advantage for such an approach to labeling aggregation.[1]

[1] Part of the results of this paper have been presented in a paper by Caminada et al. [17].

2 Preliminaries

2.1 Abstract Argumentation Framework (AAF).[2]

The seminal paper by Dung [24] introduced the fundamental notion of abstract argumentation framework that can be represented as a directed graph where the vertices represent arguments (ignoring details about their contents) and the directed arcs represent the defeat relations between these arguments.[3] For example, in Figure 1, argument A_1 is defeated by arguments A_2 and A_4 which are, in turn, defeated by arguments A_3 and A_5.

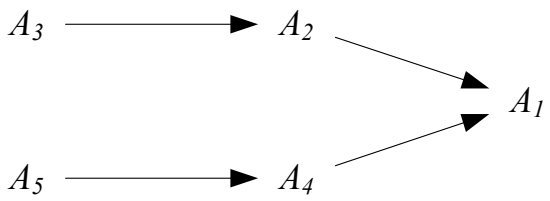

Figure 1: A simple argumentation graph

Definition 1 (Argumentation framework [24])**.** *An* argumentation framework *is a pair* $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ *where* $\mathcal{A}$ *is a finite set of arguments and* $\rightharpoonup \subseteq \mathcal{A} \times \mathcal{A}$ *is a defeat relation. We say that an argument A* defeats *an argument B if* $(A,B) \in \rightharpoonup$ *(sometimes written* $A \rightharpoonup B$*).*

There are two approaches to define semantics that assess the acceptability of arguments. One of them is extension-based semantics by Dung [24], which produces a set of arguments that are accepted together. Another equivalent labeling-based semantics is proposed by Caminada [14], which gives a labeling for each argument. With argument labelings, we can accept arguments (by labeling them as `in`), reject arguments (by labeling them as `out`), and abstain from deciding whether to accept or reject (by labeling them as `undec`). As the work by Caminada and Pigozzi [16] employed the labeling approach, so we continue to use it here.

Definition 2 (Argument labeling [14])**.** *Let* $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ *be an argumentation framework. An* argument labeling *is a total function* $\mathcal{L} : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$.

For the purposes of this paper, we use the following marking convention, as shown in Figure 2, arguments labeled `in` are shown in white, `out` in black, and `undec` in gray.

We write $\texttt{in}(\mathcal{L})$, $\texttt{out}(\mathcal{L})$, and $\texttt{undec}(\mathcal{L})$ for the set of arguments that are labeled `in`, `out`, and `undec` by $\mathcal{L}$, respectively. A labeling $\mathcal{L}$ can be represented as $\mathcal{L} = (\texttt{in}(\mathcal{L}),\texttt{out}(\mathcal{L}),\texttt{undec}(\mathcal{L}))$, or can be denote as: $\mathcal{L} = \{(A,l) |\ \mathcal{L}(A) = l \text{ for all } A \in \mathcal{A}, l \in \{\texttt{in}, \texttt{out}, \texttt{undec}\}\}$. However, labelings should follow some given conditions. If an argument is labeled `in` then all of its defeaters are labeled `out`. If an argument is labeled `out` then at least one of its defeaters is labeled `in`. We call a labeling that follows the previous two conditions an *admissible labeling*.

[2] Readers familiar with AAF can skip this part.

[3] We will use "argumentation graph" and "argumentation framework" interchangeably.

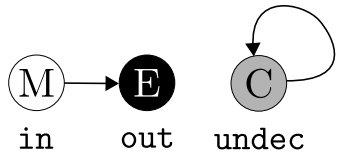

Figure 2: A labeling of an argumentation graph.

Definition 3 (Admissible labeling [14])**.** *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. An admissible labeling is a mapping $\mathcal{L} : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ such that for each $A \in \mathcal{A}$ it holds that:*
if $\mathcal{L}(A) = \texttt{in}$ then $\forall B \in \mathcal{A} : (B \rightharpoonup A \Rightarrow \mathcal{L}(B) = \texttt{out})$, and
if $\mathcal{L}(A) = \texttt{out}$ then $\exists B \in \mathcal{A} : (B \rightharpoonup A \wedge \mathcal{L}(B) = \texttt{in})$.

Some examples for *admissible* labelings, in Figure 1, can include the following: ($\{A_1, A_3, A_5\}$, $\{A_2, A_4\}$, $\emptyset$), ($\{A_3\}$, $\{A_2\}$, $\{A_1, A_4, A_5\}$), and ($\{A_5\}$, $\emptyset$, $\{A_1, A_2, A_3, A_4\}$). A *complete* labeling is an admissible labeling with the following extra condition: If an argument is labeled `undec` then there is no defeating argument that is labeled `in` and not all defeating arguments are labeled `out`.

Definition 4 (Complete labeling [14])**.** *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. A complete labeling is a mapping $\mathcal{L} : \mathcal{A} \to \{\texttt{in}, \texttt{out}, \texttt{undec}\}$ such that for each $A \in \mathcal{A}$ it holds that:*
if $\mathcal{L}(A) = \texttt{in}$ then $\forall B \in \mathcal{A} : (B \rightharpoonup A \Rightarrow \mathcal{L}(B) = \texttt{out})$,
if $\mathcal{L}(A) = \texttt{out}$ then $\exists B \in \mathcal{A} : (B \rightharpoonup A \wedge \mathcal{L}(B) = \texttt{in})$, and
if $\mathcal{L}(A) = \texttt{undec}$ then:
$\neg[\forall B \in \mathcal{A} : (B \rightharpoonup A \Rightarrow \mathcal{L}(B) = \texttt{out})] \wedge \neg[\exists B \in \mathcal{A} : (B \rightharpoonup A \wedge \mathcal{L}(B) = \texttt{in})]$

As an example for a *complete* labeling, in Figure 1, we have only one complete labeling, namely ($\{A_1, A_3, A_5\}$, $\{A_2, A_4\}$, $\emptyset$). We will use $\texttt{Adms}_{\mathcal{AF}}$ and $\texttt{Comps}_{\mathcal{AF}}$ to refer to the set of all admissible labelings and the set of all complete labelings, respectively, for an argumentation framework $\mathcal{AF}$.[4]

2.2 Aggregation Operators

Before introducing the aggregation operators that were defined by Caminada and Pigozzi [16], we first define the problem of aggregation. The problem of labeling aggregation can be formulated as a set of individuals that collectively decide how an argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ must be labelled.

Definition 5 (Labeling aggregation problem [4])**.** *Let $Ag = \{1, \ldots, n\}$ be a finite non-empty set of agents, and $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. A labeling aggregation problem is a pair $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$.*

Each individual $i \in Ag$ has a labeling $\mathcal{L}_i$ which expresses the evaluation of $\mathcal{AF}$ by this individual. A labeling profile P is a set of the labelings submitted by agents in Ag: $P = \{\mathcal{L}_1, \ldots, \mathcal{L}_n\}$.[5]

[4] $\mathcal{AF}$ will be dropped when there is no ambiguity about the argumentation framework.

[5] We follow the paper by Caminada and Pigozzi [16] in assuming that the profile is a set of labelings instead of a list of labelings. Although this is not common in judgment aggregation literature where the number of votes matter in many operators, it is not the case for the three operators considered in this study, since they focus on compatibility instead of cardinality. As such, one can think of Ag as the set of agents who submit distinct labelings for $\mathcal{AF}$, so other agents whose labelings overlap with those in P can be discarded from Ag.

A labeling aggregation operator is a function that maps a set of n labelings, chosen from the set of all labelings, $\texttt{Labs}$, into a collective labeling.[6]

Definition 6 (Labeling aggregation operator $O_{\mathcal{AF}}$ [16])**.** *Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem. A labeling aggregation operator for $\mathcal{LAP}$ is a function $O_{\mathcal{AF}} : 2^{\texttt{Labs}} \setminus \{\emptyset\} \rightarrow \texttt{Labs}$ such that $O_{\mathcal{AF}}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\}) = \mathcal{L}_{Coll}$, where $\mathcal{L}_{Coll}$ is the collective labeling.*

A labeling $\mathcal{L}_1$ is said to be *less or equally committed* as another labeling $\mathcal{L}_2$ if and only if every argument that is labeled $\texttt{in}$ by $\mathcal{L}_1$ is also labeled $\texttt{in}$ by $\mathcal{L}_2$ and every argument that is labeled $\texttt{out}$ by $\mathcal{L}_1$ is also labeled $\texttt{out}$ by $\mathcal{L}_2$.

Definition 7 (Less or equally committed $\sqsubseteq$ [16])**.** *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labelings of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We say that $\mathcal{L}_1$ is less or equally committed as $\mathcal{L}_2$ ($\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$) iff $(\texttt{in}(\mathcal{L}_1) \subseteq \texttt{in}(\mathcal{L}_2)) \wedge (\texttt{out}(\mathcal{L}_1) \subseteq \texttt{out}(\mathcal{L}_2))$.*[7]

Two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ are said to be *compatible* with each other if and only if for every argument, there is no $\texttt{in} - \texttt{out}$ conflict between the two. In other words, every argument that is labeled $\texttt{in}$ by $\mathcal{L}_1$ is not labeled $\texttt{out}$ by $\mathcal{L}_2$ and every argument that is labeled $\texttt{out}$ by $\mathcal{L}_1$ is not labeled $\texttt{in}$ by $\mathcal{L}_2$.

Definition 8 (Compatible labelings $\approx$ [16])**.** *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labelings of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We say that $\mathcal{L}_1$ is compatible with $\mathcal{L}_2$ ($\mathcal{L}_1 \approx \mathcal{L}_2$) iff $(\texttt{in}(\mathcal{L}_1) \cap \texttt{out}(\mathcal{L}_2) = \emptyset) \wedge (\texttt{out}(\mathcal{L}_1) \cap \texttt{in}(\mathcal{L}_2) = \emptyset)$*

We now define a compatible operator as the following:

Definition 9 (Compatible operator)**.** *Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem, and let $O_{\mathcal{AF}}$ be a labeling aggregation operator for $\mathcal{LAP}$. We say $O_{\mathcal{AF}}$ is a compatible operator if given any labeling profile $P = \{\mathcal{L}_1, \ldots, \mathcal{L}_n\}$, $O_{\mathcal{AF}}(P) \approx \mathcal{L}_i, \forall i \in Ag$ i.e. the outcome of $O_{\mathcal{AF}}$ is compatible with each individual's labeling.*

Caminada and Pigozzi [16] proposed three different aggregation operators, namely the skeptical operator, the credulous operator and the super credulous operator. Each of these operators maps a set of labelings, that are submitted by individuals, into a collective labeling. The following two definitions are used in the definition of these operators:

Definition 10 (Initial operators $\sqcap$, $\sqcup$ [16])**.** *Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem. The skeptical initial $\sqcap$ and credulous initial $\sqcup$ operators are labeling aggregation operators for $\mathcal{LAP}$ defined as the following:*

[6] Although it would be more precise to use $\texttt{Labs}^S_{\mathcal{AF}}$ to denote the set of all labelings for $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ according to semantics S, we will often drop $\mathcal{AF}$ and S, and use $\texttt{Labs}$ instead when there is no ambiguity about the argumentation framework. The same goes for all other notations (e.g. $O_{\mathcal{AF}}$) that were defined for an $\mathcal{AF}$, when there is no ambiguity about the argumentation framework.

[7] To improve readability, we often use the logical connectives $\wedge$, $\vee$, $\neg$, and $\Rightarrow$ instead of *and*, *or*, *not*, and *implies*, respectively.

- $\sqcap(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}) = \{(A,\texttt{in})|\forall i \in Ag : \mathcal{L}_i(A) = \texttt{in}\} \cup \{(A,\texttt{out})|\forall i \in Ag : \mathcal{L}_i(A) = \texttt{out}\} \cup \{(A,\texttt{undec})|\exists i \in Ag : \mathcal{L}_i(A) \neq \texttt{in} \wedge \exists j \in Ag : \mathcal{L}_j(A) \neq \texttt{out}\}$
- $\sqcup(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}) = \{(A,\texttt{in})|\exists i \in Ag : \mathcal{L}_i(A) = \texttt{in} \wedge \neg\exists j \in Ag : \mathcal{L}_j(A) = \texttt{out}\} \cup \{(A,\texttt{out})|\exists i \in Ag : \mathcal{L}_i(A) = \texttt{out} \wedge \neg\exists j \in Ag : \mathcal{L}_j(A) = \texttt{in}\} \cup \{(A,\texttt{undec})|\forall i \in Ag : \mathcal{L}_i(A) = \texttt{undec} \vee (\exists j \in Ag : \mathcal{L}_j(A) = \texttt{in} \wedge \exists k \in Ag : \mathcal{L}_k(A) = \texttt{out})\}$ [8]

In words, the skeptical initial operator $\sqcap$ collectively accepts (resp. rejects) an argument if it was accepted (resp. rejected) by every agent, while the credulous initial operator $\sqcup$ collectively accepts (resp. rejects) an argument if it was accepted (resp. rejected) by some agents, but not rejected (resp. accepted) by any agent. Remaining arguments are collectively labeled as undecided by both operators.

Definition 11 (Down-admissible $\downarrow$ and up-complete $\uparrow$ labelings [16])**.** *Let $\mathcal{L}$ be a labeling of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. The down-admissible labeling of $\mathcal{L}$, denoted as $\mathcal{L}\downarrow$, is the biggest (i.e. the most committed) element of the set of all admissible labelings that are less or equally committed as $\mathcal{L}$:*

$$\forall \mathcal{L}' \in \texttt{Adms} \ : \ (\mathcal{L}' \sqsubseteq \mathcal{L} \Rightarrow \mathcal{L}' \sqsubseteq (\mathcal{L}\downarrow) \sqsubseteq \mathcal{L})$$

The up-complete labeling of $\mathcal{L}$, denoted as $\mathcal{L}\uparrow$, is the smallest (i.e. the least committed) element of the set of all complete labelings that are bigger or equally committed as $\mathcal{L}$.

$$\forall \mathcal{L}' \in \texttt{Comps} \ : \ (\mathcal{L} \sqsubseteq \mathcal{L}' \Rightarrow \mathcal{L} \sqsubseteq (\mathcal{L}\uparrow) \sqsubseteq \mathcal{L}')$$

Now, we provide the definitions of the three operators:

Definition 12 (Skeptical $so_{\mathcal{AF}}$, Credulous $co_{\mathcal{AF}}$ and Super Credulous $sco_{\mathcal{AF}}$ operators [16])**.** *Let $\mathcal{LAP} = \langle Ag, \mathcal{AF} \rangle$ be a labeling aggregation problem. The skeptical $so_{\mathcal{AF}}$, the credulous $co_{\mathcal{AF}}$ and super credulous $sco_{\mathcal{AF}}$ operators are labeling aggregation operators for $\mathcal{LAP}$ defined as the following:*

- $so_{\mathcal{AF}}(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}) = (\sqcap(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}))\downarrow$.
- $co_{\mathcal{AF}}(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}) = (\sqcup(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}))\downarrow$.
- $sco_{\mathcal{AF}}(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}) = ((\sqcup(\{\mathcal{L}_1,\ldots,\mathcal{L}_n\}))\downarrow)\uparrow$.

It was shown in the work of Caminada and Pigozzi [16][Theorem 5, Theorem 11] that the down-admissible labeling of a labeling of $\mathcal{AF}$ is unique, and that the up-complete labeling of an admissible labeling of $\mathcal{AF}$ is unique. Thus, the three operators above are well defined. Further, given the set of all admissible labelings Adms for some argumentation framework, it was shown that the outcome of the skeptical aggregation operator is the biggest element in Adms that is less or equally committed as every individual's labeling.

[8] We will often use $sio_{\mathcal{AF}}$ and $cio_{\mathcal{AF}}$ to refer to the skeptical initial and credulous initial operators, respectively.

Theorem 1 ([16]). *Let $\mathcal{L}_1, \ldots, \mathcal{L}_n$ ($n \geq 1$) be labelings of argumentation framework $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. Let $\mathcal{L}_{SO} = so_{\mathcal{AF}}(\{\mathcal{L}_1, \ldots, \mathcal{L}_n\})$. It holds that $\mathcal{L}_{SO}$ is the biggest admissible labeling such that for every $i \in Ag : \mathcal{L}_{SO} \sqsubseteq \mathcal{L}_i$.*

According to Caminada and Pigozzi [16], given any profile P, the skeptical operator always produces an admissible labeling that is smaller or equally committed ($\sqsubseteq$) as each labeling in P, the credulous operator always produces an admissible labeling that is compatible ($\approx$) with each labeling in P, and the super credulous operator always produces a complete labeling that is compatible ($\approx$) with each labeling in P. Our analysis in this work aims to investigate the social optimality and strategic manipulability while respecting the properties of the outcomes of each operator. For example, a labeling that does not conform to the property of skeptical operator outcomes mentioned above (that is, admissible and is smaller or equally committed ($\sqsubseteq$) as each labeling in P, for any P) is not considered for comparison. Given this, we define here the respective set of labelings for each operator, given a profile.

Definition 13 (Respective set for skeptical $\mathcal{E}^P_{so}$, credulous $\mathcal{E}^P_{co}$, and super credulous $\mathcal{E}^P_{sco}$ operators)**.** *Given a profile P, the respective set for:*

- *skeptical operator is* $\mathcal{E}^P_{so} = \{\mathcal{L} | \mathcal{L} \in \texttt{Adms}; \mathcal{L} \sqsubseteq \mathcal{L}_i, \forall \mathcal{L}_i \in P\}$
- *credulous operator is* $\mathcal{E}^P_{co} = \{\mathcal{L} | \mathcal{L} \in \texttt{Adms}; \mathcal{L} \approx \mathcal{L}_i, \forall \mathcal{L}_i \in P\}$
- *super credulous operator is* $\mathcal{E}^P_{sco} = \{\mathcal{L} | \mathcal{L} \in \texttt{Comps}; \mathcal{L} \approx \mathcal{L}_i, \forall \mathcal{L}_i \in P\}$

Note that $\mathcal{E}^P_{so} \subseteq \mathcal{E}^P_{co}$ and $\mathcal{E}^P_{sco} \subseteq \mathcal{E}^P_{co}$, for any profile P.

2.3 Distance Measures

In this part, we define the family of distance measures that we use to define preferences. Each of the distance measures we consider is characterized by two choices:

- Set inclusion vs. Quantitative distance.
- Uniform vs. `undec` in the middle.

The combination of these choices produces four different distance measures. We start from the second choice. The uniform vs. `undec` in the middle choice captures the intuition that an `in`/`out` disagreement may be as serious or more serious (depending on the contexts) than a `in`/`undec` (or a `out`/`undec`) disagreement.

Thus, we consider the following two cases. First, `in`, `out`, and `undec` are equally distant from each other. In other words, $dist(\texttt{in}, \texttt{out}) = dist(\texttt{dec}, \texttt{undec})$, where $dist(.)$ is the difference between two labels for one argument, and `dec` is either `in` or `out`. In the other case, we assume that `undec` is in the middle between `in` and `out`. Thus, we differentiate between two types of disagreement. One between `in` and `out`, and the other between `dec` and `undec`. When considering distance, we assume $dist(\texttt{in}, \texttt{out}) > dist(\texttt{dec}, \texttt{undec})$.

2.3.1 Case 1: `in`, `out`, and `undec` are Equally Distant from Each Other

Hamming Set and Hamming Distance
The Hamming set between two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ is the set of arguments that these two labelings disagree upon.

Definition 14 (Hamming Set $\ominus$)**.** *Let $\mathcal{L}_1$, $\mathcal{L}_2$ be two labelings of $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We define the Hamming set between these two labelings as:*

$$\mathcal{L}_1 \ominus \mathcal{L}_2 = \{A \in \mathcal{A} | \mathcal{L}_1(A) \neq \mathcal{L}_2(A)\} \tag{1}$$

The Hamming distance between two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ is the number of arguments that these two labelings disagree upon.

Definition 15 (Hamming Distance $|\ominus|$)**.** *Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be two labelings of $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We define the Hamming distance between these two labelings as:*

$$\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_2 = |\mathcal{L}_1 \ominus \mathcal{L}_2| \tag{2}$$

2.3.2 Case 2: `undec` is in the Middle between `in` and `out`

In this section, we consider the case where `undec` is in the middle between `in` and `out`. Thus, we differentiate between two types of disagreement: 1) `in`/`out` disagreement, and 2) `dec`/`undec` disagreement. When considering distance, we assume $dist(\texttt{in}, \texttt{out}) = 2 \times dist(\texttt{dec}, \texttt{undec}) = 2$. [9]

To illustrate the difference from the previous case, consider the example shown in Figure 3. In this example, one can realize that the labelings $\mathcal{L}_2$ and $\mathcal{L}_3$ are equally distant from labeling $\mathcal{L}_1$ when considering Hamming set/distance.

However, one can argue that $\mathcal{L}_3$ is closer than $\mathcal{L}_2$ to $\mathcal{L}_1$. Consider the arguments in Figure 3. Labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ seem to be on completely different sides regarding their evaluations for A and B. On the other hand, the difference between $\mathcal{L}_1$ and $\mathcal{L}_3$ is less drastic, because $\mathcal{L}_3$ abstains from taking any position about A and B.

We use IUO (short for In-Undec-Out i.e. Undec is in the middle) to denote this class of preferences.

IUO Hamming Sets and IUO Hamming Distance
The $\texttt{in} - \texttt{out}$ Hamming set ($\ominus^{io}$) between two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ is the set of arguments that both labelings label as decided (i.e. `in` or `out`), but on which they disagree upon. The $\texttt{dec} - \texttt{undec}$ Hamming set ($\ominus^{du}$) between two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ is the set of arguments that one of the two labelings labels as decided (whether `in` or `out`) and the other labels as undecided.

[9] The use of 2 here is chosen carefully to satisfy the triangle inequality. Otherwise $dist(\texttt{in}, \texttt{out})$ would not be a valid measure of the shortest distance between `in` and `out`. However, the use of any α s.t. $1 < \alpha \leq 2$ would not affect the results of this paper. We just use 2 here for simplicity.

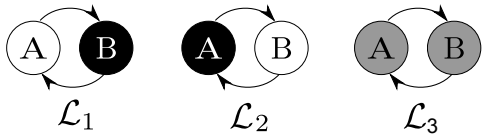

Figure 3: An example showing the need for considering `undec` as a middle labeling between `in` and `out`.

Definition 16 (IUO Hamming sets $\ominus^{\mathcal{M}}$)**.** *Let $\mathcal{L}_1$, $\mathcal{L}_2$ be two labelings of $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We define the IUO Hamming sets as a pair $\ominus^{\mathcal{M}} = (\ominus^{io}, \ominus^{du})$, where $\ominus^{io}$ is* $\texttt{in} - \texttt{out}$ *Hamming set and $\ominus^{du}$ is* $\texttt{dec} - \texttt{undec}$ *Hamming set:*

$$\mathcal{L}_1 \ominus^{io} \mathcal{L}_2 = \{A \in \mathcal{A} | (\mathcal{L}_1(A) = \texttt{in} \wedge \mathcal{L}_2(A) = \texttt{out}) \vee (\mathcal{L}_1(A) = \texttt{out} \wedge \mathcal{L}_2(A) = \texttt{in})\} \tag{3}$$

$$\mathcal{L}_1 \ominus^{du} \mathcal{L}_2 = \{A \in \mathcal{A} | (A \in \texttt{dec}(\mathcal{L}_1) \wedge \mathcal{L}_2(A) = \texttt{undec}) \vee (\mathcal{L}_1(A) = \texttt{undec} \wedge A \in \texttt{dec}(\mathcal{L}_2))\} \tag{4}$$

where $\texttt{dec}(\mathcal{L}_1)$ *is the set of decided (*`in` *or* `out`*) arguments according to the labeling $\mathcal{L}_1$.*

The IUO Hamming distance between two labelings $\mathcal{L}_1$ and $\mathcal{L}_2$ is the number of arguments in $\mathcal{L}_1 \ominus^{du} \mathcal{L}_2$ added to twice the number of the arguments in $\mathcal{L}_1 \ominus^{io} \mathcal{L}_2$.

Definition 17 (IUO Hamming Distance $\left|\ominus^{\mathcal{M}}\right|$)**.** *Let $\mathcal{L}_1$, $\mathcal{L}_2$ be two labelings of $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$. We define the IUO Hamming distance between these two labelings as:*

$$\mathcal{L}_1 \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_2 = 2 \times |\mathcal{L}_1 \ominus^{io} \mathcal{L}_2| + |\mathcal{L}_1 \ominus^{du} \mathcal{L}_2| \tag{5}$$

Table 1 summarizes the distance measures we consider.

		Uniform	**IUO**
Hamming	**Set**	Hamming Set $\ominus$	IUO Hamming Sets $\ominus^{\mathcal{M}}$
	Distance	Hamming Distance $\lvert\ominus\rvert$	IUO Hamming Distance $\left\lvert\ominus^{\mathcal{M}}\right\rvert$

Table 1: Full family of distance measures.

2.4 Preferences

Given the distance measures defined earlier, we define agents' preferences. We say an agent's preferences are *x*-based, if her preferences are calculated using the distance measure *x* (e.g. Hamming distance based preferences). We use $\succeq_{i,x}$ to denote a *weak preference* relation by agent *i*

whose preferences are x-based i.e. for any pair $\mathcal{L}_1, \mathcal{L}_2 \in \texttt{Labs}$, $\mathcal{L}_1 \succeq_{i,x} \mathcal{L}_2$ denotes that $\mathcal{L}_1$ is more or equally preferred than $\mathcal{L}_2$ by agent i with x-based preferences. Further, we use $\succ_{i,x}$ to denote a *strict preference* relation ($\mathcal{L}_1 \succ_{i,x} \mathcal{L}_2$ iff $(\mathcal{L}_1 \succeq_{i,x} \mathcal{L}_2) \wedge \neg(\mathcal{L}_2 \succeq_{i,x} \mathcal{L}_1)$), and $\sim$ to denote an *indifference* relation ($\mathcal{L}_1 \sim_{i,x} \mathcal{L}_2$ iff $(\mathcal{L}_1 \succeq_{i,x} \mathcal{L}_2) \wedge (\mathcal{L}_2 \succeq_{i,x} \mathcal{L}_1)$).

We define the subset relation over pairs of sets as the following.

Definition 18 (Subset Over Pairs $\subseteq$)**.** *Let A_1, A_2, B_1, B_2 be four sets, and Let $S_1 = (A_1, B_1)$, $S_2 = (A_2, B_2)$ be two pairs of sets. We use $S_1 \subseteq S_2$ to denote the subset relation over pairs of subsets:*

$$S_1 \subseteq S_2 \text{ iff } A_1 \subseteq A_2 \wedge B_1 \subseteq B_2 \tag{6}$$

Given a set measure $\otimes \in \{\ominus, \ominus^{\mathcal{M}}\}$, an agent i, who has $\otimes$-set based preferences (and whose top preference is $\mathcal{L}_i$), would prefer a labeling $\mathcal{L}$ over another labeling $\mathcal{L}'$ if and only if the set of arguments in $\mathcal{L}_i \otimes \mathcal{L}$ is a subset of $\mathcal{L}_i \otimes \mathcal{L}'$ (where "subset" here refers to the standard definition of subset as well as the definition of "subset over pairs" defined above). Note that the set based preference yields a partial order over the labelings.[10]

Definition 19 (Set Based Preference $\succeq_{i,\otimes}$)**.** *We say that agent i's preferences are $\otimes$-set based w.r.t $\mathcal{L}_i$ iff:*

$$\forall \mathcal{L}, \mathcal{L}' \in \texttt{Labs} : \mathcal{L} \succeq_{i,\otimes} \mathcal{L}' \Leftrightarrow \mathcal{L} \otimes \mathcal{L}_i \subseteq \mathcal{L}' \otimes \mathcal{L}_i \tag{7}$$

where $\mathcal{L}_i$ is agent i's most preferred labeling and $\otimes \in \{\ominus, \ominus^{\mathcal{M}}\}$. Note that $\otimes$-set based preferences is read Hamming set based preferences when $\otimes = \ominus$, ...etc.

Given a distance measure $|\otimes| \in \{|\ominus|, |\ominus^{\mathcal{M}}|\}$, an agent i, who has $|\otimes|$-distance based preferences (and whose top preference is $\mathcal{L}_i$), would prefer a labeling $\mathcal{L}$ over another labeling $\mathcal{L}'$ if and only if $\mathcal{L}_i |\otimes| \mathcal{L}$ is less than $\mathcal{L}_i |\otimes| \mathcal{L}'$. Note that the distance based preference yields a total pre-order over the labelings.

We now define the classes of preferences which are based on different distance measures, that we defined earlier.

Definition 20 (Distance Based Preference $\succeq_{i,|\otimes|}$)**.** *We say that agent i's preferences are $|\otimes|$-distance based w.r.t $\mathcal{L}_i$ iff:*

$$\forall \mathcal{L}, \mathcal{L}' \in \texttt{Labs} : \mathcal{L} \succeq_{i,|\otimes|} \mathcal{L}' \Leftrightarrow \mathcal{L} |\otimes| \mathcal{L}_i \leq \mathcal{L}' |\otimes| \mathcal{L}_i \tag{8}$$

where $\mathcal{L}_i$ is agent i's most preferred labeling and $|\otimes| \in \{|\ominus|, |\ominus^{\mathcal{M}}|\}$. Note that $|\otimes|$-distance based preferences is read Hamming distance based preferences when $|\otimes| = |\ominus|$, ...etc.

Note here that Hamming distance was used by Dietrich and List [21] to define preferences over sets of accepted/rejected issues in the judgment aggregation (JA) domain. Additionally, in the same work, they defined *closeness-respecting* preferences, which correspond to Hamming set based preferences in JA.

[10] Although formally, the set-based criteria are not measures but mappings to sets, we will slightly abuse terminology and refer to all criteria (set based and distance based) as *set and distance measures* for easy reference.

To illustrate the set and distance based preferences, we use Hamming set and Hamming distance based preferences for their simplicity. Consider the example in Figure 4 with four possible complete labelings. The Hamming sets between $\mathcal{L}_1$ and the other three labelings are:

$$\mathcal{L}_1 \ominus \mathcal{L}_2 = \{A, B\}$$

$$\mathcal{L}_1 \ominus \mathcal{L}_3 = \{C, D, E\}$$

$$\mathcal{L}_1 \ominus \mathcal{L}_4 = \{A, B, C, D, E\}$$

Consequently, the Hamming distance values between $\mathcal{L}_1$ and the other three labelings are the cardinality values of the Hamming sets between $\mathcal{L}_1$ and the other three labelings.

$$\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_2 = |\{A, B\}| = 2$$

$$\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_3 = |\{C, D, E\}| = 3$$

$$\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_4 = |\{A, B, C, D, E\}| = 5$$

Assume we have agents with Hamming set based preferences. Hence, an arbitrary agent i who prefers $\mathcal{L}_1$ the most, would have the following preferences: $\mathcal{L}_1 \succ \mathcal{L}_2 \succ \mathcal{L}_4$ and $\mathcal{L}_1 \succ \mathcal{L}_3 \succ \mathcal{L}_4$ (neither $\mathcal{L}_1 \ominus \mathcal{L}_2$ nor $\mathcal{L}_1 \ominus \mathcal{L}_3$ is a subset of the other). However, if agents have Hamming distance based preferences, an agent who prefers $\mathcal{L}_1$ the most, would have the following preferences: $\mathcal{L}_1 \succ \mathcal{L}_2 \succ \mathcal{L}_3 \succ \mathcal{L}_4$.

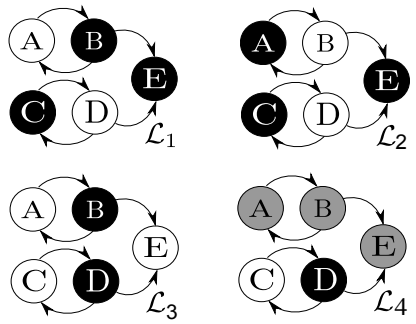

Figure 4: An argumentation graph with four possible complete labelings.

The following lemma is important in the context of compatible operators. For each agent $i \in Ag$, let $\mathcal{L}_i = \mathcal{L}_1$. Then, provided the conditions below, the lemma says an individual's preference over $\mathcal{L}_2$ and $\mathcal{L}_3$ would coincide whether she has a Hamming set (resp. distance) or IUO Hamming sets (resp. distance).

Lemma 1. *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Let $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ be three labelings and let $\mathcal{L}_1 \approx \mathcal{L}_2$ and $\mathcal{L}_1 \approx \mathcal{L}_3$:*

$\mathcal{L}_1 \ominus \mathcal{L}_2 \subseteq \mathcal{L}_1 \ominus \mathcal{L}_3 \Leftrightarrow \mathcal{L}_1 \ominus^{\mathcal{M}} \mathcal{L}_2 \subseteq \mathcal{L}_1 \ominus^{\mathcal{M}} \mathcal{L}_3$, and
$\mathcal{L}_1 \,|\ominus|\, \mathcal{L}_2 \leq \mathcal{L}_1 \,|\ominus|\, \mathcal{L}_3 \Leftrightarrow \mathcal{L}_1 \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_2 \leq \mathcal{L}_1 \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_3$.

(or equivalently $\mathcal{L}_2 \succeq_{1,\ominus} \mathcal{L}_3 \Leftrightarrow \mathcal{L}_2 \succeq_{1,\ominus^{\mathcal{M}}} \mathcal{L}_3$ and $\mathcal{L}_2 \succeq_{1,|\ominus|} \mathcal{L}_3 \Leftrightarrow \mathcal{L}_2 \succeq_{1,|\ominus^{\mathcal{M}}|} \mathcal{L}_3$)

Proof sketch. Since $\mathcal{L}_1 \approx \mathcal{L}_2$, there is no in-out disagreement between them (i.e. $\mathcal{L}_1 \ominus^{io} \mathcal{L}_2 = \emptyset$). This makes the Hamming set and the IUO Hamming set between $\mathcal{L}_1$ and $\mathcal{L}_2$ equivalent. Consequently, the Hamming distance and the IUO Hamming distance between them are also equivalent. The same goes for $\mathcal{L}_1$ and $\mathcal{L}_3$ (since $\mathcal{L}_1 \approx \mathcal{L}_3$). The rest follows from the definition of the set-based and the distance-based preferences. ■

The following lemmas are also crucial for the proofs of theorems in this paper. Since the labelings have only three values, we can use the following lemma.

Lemma 2. *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Let* $\mathtt{dec}(\mathcal{L}) = \mathtt{in}(\mathcal{L}) \cup \mathtt{out}(\mathcal{L})$ $\forall \mathcal{L} \in \mathtt{Labs}$. *For any pair $\mathcal{L}_1, \mathcal{L}_2 \in \mathtt{Labs}$:*

a) $\mathcal{L}_1 \ominus \mathcal{L}_2 = (\mathtt{in}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2)) \cup (\mathtt{in}(\mathcal{L}_1) \cap \mathtt{undec}(\mathcal{L}_2)) \cup (\mathtt{out}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2)) \cup (\mathtt{out}(\mathcal{L}_1) \cap \mathtt{undec}(\mathcal{L}_2)) \cup (\mathtt{undec}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2)) \cup (\mathtt{undec}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2))$

b) if $\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$ *then* $\mathcal{L}_1 \ominus \mathcal{L}_2 = \mathtt{undec}(\mathcal{L}_1) \cap \mathtt{dec}(\mathcal{L}_2)$

c) if $\mathcal{L}_1 \approx \mathcal{L}_2$ *then* $\mathcal{L}_1 \ominus \mathcal{L}_2 = (\mathtt{dec}(\mathcal{L}_1) \cap \mathtt{undec}(\mathcal{L}_2)) \cup (\mathtt{undec}(\mathcal{L}_1) \cap \mathtt{dec}(\mathcal{L}_2))$

Proof.

a) This follows from the fact that $\mathtt{in}(\mathcal{L})$, $\mathtt{out}(\mathcal{L})$ and $\mathtt{undec}(\mathcal{L})$ partition the domain of any labeling $\mathcal{L}$.

b) From $\mathcal{L}_1 \sqsubseteq \mathcal{L}_2$, the sets $(\mathtt{in}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2))$, $(\mathtt{in}(\mathcal{L}_1) \cap \mathtt{undec}(\mathcal{L}_2))$, $(\mathtt{out}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2))$, and $(\mathtt{out}(\mathcal{L}_1) \cap \mathtt{undec}(\mathcal{L}_2))$ are all empty sets. Then, we are left with the following:

$$(\mathtt{undec}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2)) \cup (\mathtt{undec}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2))$$

which can be written as:

$$\mathtt{undec}(\mathcal{L}_1) \cap (\mathtt{in}(\mathcal{L}_2) \cup \mathtt{out}(\mathcal{L}_2))$$

and replacing $\mathtt{in}(\mathcal{L}) \cup \mathtt{out}(\mathcal{L})$ by $\mathtt{dec}(\mathcal{L})$ would give the result.

c) From $\mathcal{L}_1 \approx \mathcal{L}_2$, the sets $(\mathtt{in}(\mathcal{L}_1) \cap \mathtt{out}(\mathcal{L}_2))$, and $(\mathtt{out}(\mathcal{L}_1) \cap \mathtt{in}(\mathcal{L}_2))$ are empty. The rest can be rearranged similarly to b), and replacing $\mathtt{in}(\mathcal{L}) \cup \mathtt{out}(\mathcal{L})$ by $\mathtt{dec}(\mathcal{L})$ would give the result.

■

We now prove two lemmas establishing the relations between less or equally committed labelings and Hamming based preferences over labelings.

Lemma 3. *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Let $\mathcal{L}$, $\mathcal{L}'$ and $\mathcal{L}_i$ be three labelings such that $\mathcal{L} \sqsubseteq \mathcal{L}' \sqsubseteq \mathcal{L}_i$. If $\mathcal{L}_i$ is the most preferred labeling of agent i and her preference is Hamming set based or Hamming distance based, then $\mathcal{L}' \succeq_{i,\ominus} \mathcal{L}$ and $\mathcal{L}' \succeq_{i,|\ominus|} \mathcal{L}$ respectively.*

Proof. From $\mathcal{L} \sqsubseteq \mathcal{L}'$, we have that $\texttt{dec}(\mathcal{L}) \subseteq \texttt{dec}(\mathcal{L}')$, which is equivalent to $\texttt{undec}(\mathcal{L}') \subseteq \texttt{undec}(\mathcal{L})$ because undec is the complement of dec. From this, it follows that $\texttt{undec}(\mathcal{L}') \cap \texttt{dec}(\mathcal{L}_i) \subseteq \texttt{undec}(\mathcal{L}) \cap \texttt{dec}(\mathcal{L}_i)$. Since $\mathcal{L} \sqsubseteq \mathcal{L}_i$ and $\mathcal{L}' \sqsubseteq \mathcal{L}_i$ (by assumption and transitivity of $\sqsubseteq$), we can use Lemma 2b to obtain $\mathcal{L}' \ominus \mathcal{L}_i \subseteq \mathcal{L} \ominus \mathcal{L}_i$. By definition we have that $\mathcal{L}' \succeq_{i,\ominus} \mathcal{L}$ and $\mathcal{L}' \succeq_{i,|\ominus|} \mathcal{L}$. ■

Lemma 4. *Let $\mathcal{AF} = \langle \mathcal{A}, \rightharpoonup \rangle$ be an argumentation framework. Let $\mathcal{L}$, $\mathcal{L}'$ and $\mathcal{L}_i$ be three labelings and let $\mathcal{L} \sqsubseteq \mathcal{L}_i$. If $\mathcal{L}_i$ is the most preferred labeling of agent i, her preference is Hamming set based and $\mathcal{L}' \succeq_{i,\ominus} \mathcal{L}$, then $\mathcal{L} \sqsubseteq \mathcal{L}'$.*

Proof. $\mathcal{L}' \succeq_{i,\ominus} \mathcal{L}$ implies $\mathcal{L}' \ominus \mathcal{L}_i \subseteq \mathcal{L} \ominus \mathcal{L}_i$ which implies $\mathcal{L}(A) = \mathcal{L}_i(A) \Rightarrow \mathcal{L}'(A) = \mathcal{L}_i(A)$ for any argument A (i). Now, $\mathcal{L} \sqsubseteq \mathcal{L}_i$ implies $\mathcal{L}(A) = \mathcal{L}_i(A)$ for any $A \in \texttt{dec}(\mathcal{L})$ (ii). From (i) and (ii) it follows that $\mathcal{L}(A) = \mathcal{L}'(A)$ for any $A \in \texttt{dec}(\mathcal{L})$. Hence $\mathcal{L} \sqsubseteq \mathcal{L}'$. ■

3 Pareto Optimality

In this section, we study the Pareto optimality of the outcomes of the three operators given different variations of the preferences. Pareto optimality is one of the fundamental concepts that ensures that, given a profile, the social outcome selected by the aggregation procedure cannot be improved.

A labeling $\mathcal{L}_1$ Pareto dominates $\mathcal{L}_2$ if and only if for any agent i, i would prefer $\mathcal{L}_1$ at least as much as she prefers $\mathcal{L}_2$, and for at least one agent j, j would strictly prefer $\mathcal{L}_1$ over $\mathcal{L}_2$.

Definition 21 (Pareto dominance)**.** *Let $Ag = \{1, \ldots, n\}$ be a set of agents with preferences $\succeq_i$, $i \in Ag$. $\mathcal{L}$ Pareto dominates $\mathcal{L}'$ iff $\forall i \in Ag$, $\mathcal{L} \succeq_i \mathcal{L}'$ and $\exists j \in Ag$, $\mathcal{L} \succ_j \mathcal{L}'$.*

A labeling is Pareto optimal in a set, if it is not Pareto dominated by any other labeling from that set.

Definition 22 (Pareto optimality of a labeling in $\mathbb{S}$)**.** *Let $\mathbb{S}$ be a set of labelings. A labeling $\mathcal{L}$ is Pareto optimal in $\mathbb{S}$ if there is no labeling $\mathcal{L}' \in \mathbb{S}$ such that $\mathcal{L}'$ Pareto dominates $\mathcal{L}$.*

In our results, the set $\mathbb{S}$ will mainly refer to one of the three respective sets (from Def.13). Moreover, whenever we refer to an operator as Pareto optimal (in a set $\mathbb{S}$) we mean that it only produces Pareto optimal outcomes (in $\mathbb{S}$).

Definition 23 (Pareto optimality of an operator in $\mathbb{S}$)**.** *Let $\mathbb{S}$ be a set of labelings. An operator is Pareto optimal in $\mathbb{S}$ if it only produces Pareto optimal (in $\mathbb{S}$) outcomes.*

Similarly, the set S will mainly refer to the respective set of the operator, in the case of the three operators.

3.1 Connections between Classes of Preferences

We notice that Pareto optimality carries over from each of the distance-based preferences to its corresponding set-based preferences.

Proposition 1. *Let $\otimes \in \{\ominus, \ominus^{\mathcal{M}}\}$ be a set measure and $|\otimes|$ be its corresponding distance measure (i.e. if $\otimes = \ominus^{\mathcal{M}}$ then $|\otimes| = |\ominus^{\mathcal{M}}|$). If a labeling[11] is Pareto optimal in a set $\mathcal{S}$ given agents with $|\otimes|$-based preferences, then it is Pareto optimal in $\mathcal{S}$ given agents with $\otimes$-based preferences.*

Proof sketch. Suppose, towards a contradiction, that there exists a labeling $\mathcal{L}$ that is Pareto optimal in a set $\mathcal{S}$ given agents with $|\otimes|$-based preferences, but $\mathcal{L}$ is not Pareto optimal in $\mathcal{S}$ given agents with $\otimes$-based preferences. Then, there exists a labeling $\mathcal{L}_X$ in $\mathcal{S}$ such that $\mathcal{L}_X$ Pareto dominates $\mathcal{L}$ in S (given agents with $\otimes$-based preferences). Then, for each agent i, whenever $\mathcal{L}_X$ disagrees with $\mathcal{L}_i$ on an argument, $\mathcal{L}$ would also disagree in the same way with $\mathcal{L}_i$ on the same argument. Further, for at least one agent j, $\mathcal{L}$ would disagree with $\mathcal{L}_j$ on some argument for which $\mathcal{L}_X$ agrees with $\mathcal{L}_j$ on. Since for any two sets $A, B : (A \subseteq B \Rightarrow |A| \leq |B|)$ (for both standard subsets and subsets over pairs, as defined in Def.18), then $\mathcal{L}_X$ would also Pareto dominate $\mathcal{L}$ in S given agents with $|\otimes|$-based preferences. Contradiction. ■

While these connections are only one-way, they hold without restrictions. However, when further restrictions are introduced, one can find more connections. The following result shows that when all labelings in $\mathcal{S}$ are admissible labelings and are compatible ($\approx$) with each of the individuals' labelings, some other connections hold.

Proposition 2. *Let $\mathcal{S}$ be any arbitrary set such that $\mathcal{S} \subseteq \mathcal{E}^P_{co}$, for an arbitrary P. A labeling from $\mathcal{S}$ is Pareto optimal in $\mathcal{S}$ when individual preferences are Hamming set (resp. distance) based iff it is Pareto optimal in $\mathcal{S}$ when individual preferences are IUO Hamming sets (resp. distance) based.*

Proof sketch. Since all labelings in S are compatible with every individual's labeling, using Lemma 1, Hamming set (resp. distance) based preference and IUO Hamming set (resp. distance) based preferences would be equivalent for each agent. Thus, the preference order would be the same for each agent whether she is using Hamming set (resp. distanced) based preferences, or IUO Hamming set (resp. distanced) based preferences. ■

Note that these connections hold in both directions, unlike in the previous result where connections are one-way (from distance based to set based, but not vice versa). Other than the ones found above, there exist no more connections, even after considering further restrictions, similar to the ones in the previous part. One can provide counterexamples for the connections that do not hold between the classes of preferences. We summarize all the findings in Table 2. Now we turn to studying the Pareto optimality of the three operators; the skeptical, the credulous and the super credulous, with respect to the four classes of preferences.

3.2 Hamming Set and Hamming Distance

In this part, we establish the first advantage of the skeptical operator over the credulous and super credulous operators. When all individuals' preferences are Hamming set based, or all are Hamming distance based, the skeptical operator is Pareto optimal in the set of admissible labelings that are smaller or equally committed ($\sqsubseteq$) as each individual's labeling.

[11] Note that since an operator is Pareto optimal in a set if and only if all of its outcomes are Pareto optimal in that set, then one can see that in this theorem, and others as well, 'labeling' can be substituted with 'operator'.

	HS	HD	IUO HS	IUO HD
Hamming set (HS)	Y	N	Y*	N
Hamming dist. (HD)	Y	Y	Y*	Y*
IUO Hamming sets (IUO HS)	Y*	N	Y	N
IUO Hamming dist. (IUO HD)	Y*	Y*	Y	Y

Table 2: Pareto optimality relations between the different preference classes. A Y means Pareto optimality carries over from the class in the row to the class in the column, a Y^* means it only carries over if the operator only produces compatible labelings, and an N means that it does not necessarily carry over even if the operator only produces compatible labelings.

Theorem 2. *If individual preferences are Hamming distance based, then the skeptical aggregation operator is Pareto optimal in its respective set.*

Proof. Let P be a profile of labelings, $\mathcal{L}_{SO} = so_{\mathcal{AF}}(P)$ and $\mathcal{L}_X$ be some admissible labeling with the property $\forall i \in Ag, \mathcal{L}_X \sqsubseteq \mathcal{L}_i$. From Theorem 1 we know that $\mathcal{L}_{SO}$ is the biggest admissible labeling with such property, so $\mathcal{L}_X \sqsubseteq \mathcal{L}_{SO}$. So we have $\forall i \in Ag, \mathcal{L}_X \sqsubseteq \mathcal{L}_{SO} \sqsubseteq \mathcal{L}_i$. From Lemma 3 we have $\mathcal{L}_{SO} \succeq_{i,|\ominus|} \mathcal{L}_X$ for any i. So no agent strictly prefers $\mathcal{L}_X$ and hence there is no labeling that Pareto dominates $\mathcal{L}_{SO}$. ■

Corollary 1. *If individual preferences are Hamming set based, then the skeptical aggregation operator is Pareto optimal in its respective set.*

Proof. From Theorem 2 and Proposition 1. ■

On the other hand, the credulous and super credulous operators are only Pareto optimal when individuals have Hamming set based preferences, and they fail to produce Pareto optimal outcomes when the preferences are Hamming distance based. The proof for the following theorem is omitted due to its similarity to the one for Theorem 4. The interested reader can see the full proof in the work by Caminada et al. [17].

Theorem 3. *If individual preferences are Hamming set based, then the credulous aggregation operator is Pareto optimal in its respective set.*

Theorem 4. *If individual preferences are Hamming set based, then the super credulous aggregation operator is Pareto optimal in its respective set.*

Proof. Let P be a profile of labelings, $\mathcal{L}_{CIO} = \sqcup(P)$, $\mathcal{L}_{CO} = co_{\mathcal{AF}}(P)$, and $\mathcal{L}_{SCO} = sco_{\mathcal{AF}}(P)$. Suppose, towards a contradiction, that there exists a complete labeling $\mathcal{L}_X$ s.t. $\mathcal{L}_X \approx \mathcal{L}_i \ \forall i \in Ag$, and $\mathcal{L}_X$ dominates $\mathcal{L}_{SCO}$ (w.r.t $\succeq_{i,\ominus}$).

Let $A \in \texttt{dec}(\mathcal{L}_{CO})$, then $\mathcal{L}_{SCO}$ agrees on A with $\mathcal{L}_{CO}$. However, $\mathcal{L}_{CO}$ only decides on an argument if at least one agent decides on this argument and agrees with $\mathcal{L}_{CO}$ on it. Then, this agent also agrees on A with $\mathcal{L}_{SCO}$. Since $\mathcal{L}_X$, by assumption, Pareto dominates $\mathcal{L}_{SCO}$, $\mathcal{L}_X$ also needs to agree with this agent on A. This is the case for every argument $A \in \texttt{dec}(\mathcal{L}_{CO})$. Hence, $\forall A \in \texttt{dec}(\mathcal{L}_{CO}) : \mathcal{L}_{CO}(A) = \mathcal{L}_X(A)$. Then, $\mathcal{L}_{CO} \sqsubseteq \mathcal{L}_X$. By definition, $\mathcal{L}_{SCO}$ is the smallest element

(w.r.t $\sqsubseteq$) of the set of all complete labelings that are bigger or equally committed as $\mathcal{L}_{CO}$. Then, $\mathcal{L}_{CO} \sqsubseteq \mathcal{L}_{SCO} \sqsubseteq \mathcal{L}_X$.

$\mathcal{L}_X$ should be different from $\mathcal{L}_{SCO}$ to dominate it. Then, $\exists A \in \texttt{undec}(\mathcal{L}_{SCO}) \cap \texttt{dec}(\mathcal{L}_X)$. We will show that $\forall A \in \texttt{undec}(\mathcal{L}_{SCO}) \cap \texttt{dec}(\mathcal{L}_X)$ then $\forall i \in Ag : \mathcal{L}_i(A) = \texttt{undec}$. This is enough to reach a contradiction because it shows that all agents agree on at least one argument with $\mathcal{L}_{SCO}$ while disagree with $\mathcal{L}_X$ on that argument.

Suppose, for contradiction, that $\exists A \in \texttt{undec}(\mathcal{L}_{SCO}) \cap \texttt{dec}(\mathcal{L}_X)$, and there exists an agent j such that $\mathcal{L}_j(A) = \mathcal{L}_X(A) \in \{\texttt{in}, \texttt{out}\}$. Since $A \in \texttt{undec}(\mathcal{L}_{SCO})$, then $A \in \texttt{undec}(\mathcal{L}_{CO})$. However, $\mathcal{L}_X$ is a complete labeling which means that it is also an admissible labeling, and from Theorem 3, $\mathcal{L}_{CO}$ is Pareto optimal in the set of all admissible labelings that are compatible ($\approx$) with each of the participants' labelings (i.e. the respective set of the credulous operator). Then:

$$\forall B \in \mathcal{A}, \neg \exists i \in Ag \text{ s.t. } \mathcal{L}_{CO}(B) \neq \mathcal{L}_i(B) \wedge \mathcal{L}_X(B) = \mathcal{L}_i(B) \tag{9}$$

Contradiction. Then, all agents need to agree with $\mathcal{L}_{CO}$ and $\mathcal{L}_{SCO}$ on every A s.t. $A \in \texttt{undec}(\mathcal{L}_{SCO}) \cap \texttt{dec}(\mathcal{L}_X)$ (and disagree with $\mathcal{L}_X$ on A). ■

Observation 1. *If individual preferences are Hamming distance based, then neither the credulous nor the super credulous aggregation operator is Pareto optimal in their respective sets. An example is given in Figure 5 where $\mathcal{L}_{CO}$ represents the outcome of the credulous (and the super credulous) aggregation operator. Both labelings $\mathcal{L}_{CO}$ and $\mathcal{L}_X$ are compatible with both $\mathcal{L}_1$ and $\mathcal{L}_2$, but $\mathcal{L}_X$ is closer when applying Hamming distance. $\mathcal{L}_1 \ominus \mathcal{L}_{CO} = \mathcal{L}_2 \ominus \mathcal{L}_{CO} = \{A, B, E, F, G\}$, so the Hamming distance is 5, whereas $\mathcal{L}_1 \ominus \mathcal{L}_X = \mathcal{L}_2 \ominus \mathcal{L}_X = \{A, B, C, D\}$, so the Hamming distance is 4.*

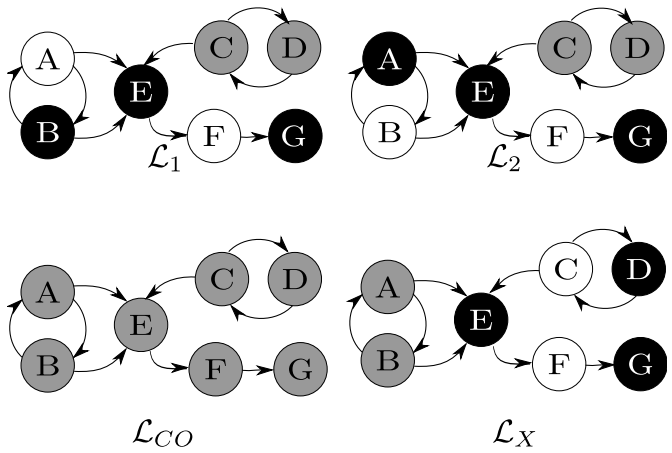

Figure 5: If individuals' preferences are Hamming distance based, the (super) credulous aggregation operator is not Pareto optimal in its respective set.

3.3 IUO Hamming Sets and IUO Hamming Distance

We now analyze the Pareto optimality for the three operators given the classes of preferences that assume undec to be in the middle between in and out ($dist(\texttt{dec},\texttt{undec}) < dist(\texttt{in},\texttt{out})$). We use Proposition 2 to show that the results for the three operators given IUO Hamming set and distance based preferences echo their results with the Hamming set and distance based preferences.

Corollary 2. *If individual preferences are IUO Hamming sets based, then the skeptical, the credulous, and the super credulous aggregation operator are Pareto optimal in their respective sets.*

Proof. From Proposition 2 with each of Corollary 1, Theorem 3, and Theorem 4, respectively. ■

Corollary 3. *If individual preferences are IUO Hamming distance based, then the skeptical aggregation operator is Pareto optimal in its respective set, but the credulous, and the super credulous aggregation operator are not Pareto optimal in their respective sets.*

Proof. From Proposition 2 and Theorem 2 for the skeptical operator; and from Proposition 2 and Observation 1 for the credulous and the super credulous operators. ■

Table 3 summarizes the Pareto optimality results for the three operators given all the eight classes of preferences.

	Skeptical Operator	**Credulous Operator**	**Super Credulous Operator**
Hamming set	Yes (Cor. 1)	Yes (Thm. 3)	Yes (Thm. 4)
Hamming dist.	Yes (Thm. 2)	No (Obs. 1)	No (Obs. 1)
IUO Hamming sets	Yes (Cor. 2)	Yes (Cor. 2)	Yes (Cor. 2)
IUO Hamming dist.	Yes (Cor. 3)	No (Cor. 3)	No (Cor. 3)

Table 3: Pareto optimality in the respective set of the aggregation operators depending on the type of preference.

3.4 Heterogeneous Preferences

The previous subsections have considered the case where agents have homogeneous preferences i.e. agents share the same class of preferences (e.g. all agents have Hamming set based preferences). However, there can be some scenarios where this assumption does not hold. In this part, we study the effect of removing this assumption.

Let $\mathcal{F}$ be the set of all classes of preferences, $\mathcal{R}$ be some arbitrary subset of $\mathcal{F}$, and $c : Ag \to \mathcal{F}$ be a function defining the class of preferences for each agent. We say that the set of agents Ag have *homogeneous* preferences from $\mathcal{R}$ if $\forall i, j \in Ag : c(i) = c(j) \in \mathcal{R}$. We say Ag have *heterogeneous* preferences from $\mathcal{R}$ if $\forall i \in Ag : c(i) \in \mathcal{R}$ and $\exists i, j \in Ag$ s.t. $c(i) \neq c(j)$.

Let $\mathcal{R}$ be an arbitrary set of classes of preferences. In general, if a labeling $\mathcal{L}$ is Pareto optimal in a set $\mathcal{S}$ given that *Ag* have *homogeneous* preferences from $\mathcal{R}$, then $\mathcal{L}$ might not be Pareto optimal if *Ag* have *heterogeneous* preferences from $\mathcal{R}$.

However, one can show that some of the classes of preferences that we defined enjoy special relations with each others that make Pareto optimality carry over from homogeneous preference of each of those classes to heterogeneous preferences that combine all of those classes. Consider the following theorem.

Theorem 5. *Let $\mathcal{R} = \{\ominus, \ominus^{\mathcal{M}}\}$ be a set of preference classes, Ag be a set of agents, and $\mathcal{L}$ be a labeling from $\mathcal{E}^{P}_{co}$ (the respective set of the credulous operator, for an arbitrary P). If $\mathcal{L}$ is Pareto optimal in $\mathcal{E}^{P}_{co}$ given that Ag have* homogeneous *preferences from $\mathcal{R}$, then $\mathcal{L}$ is Pareto optimal in $\mathcal{E}^{P}_{co}$ given that Ag have* heterogeneous *preferences from $\mathcal{R}$.*

Proof sketch. Given the compatibility of all labelings from $\mathcal{E}^{P}_{co}$ with every individuals' labeling, and from Lemma 1, if some agents who have Hamming set based preferences switched their classes of preferences to IUO Hamming sets based preferences or vice versa, then their preferences would not change. ■

For our three operators, we have the following corollary.

Corollary 4. *Let $\mathcal{R} = \{\ominus, \ominus^{\mathcal{M}}\}$. The skeptical, the credulous, and the super credulous aggregation operators are Pareto optimal in their respective sets given that individuals have heterogeneous preferences from $\mathcal{R}$.*

We showed earlier that the skeptical operator is always Pareto optimal no matter which class of preferences the individuals have, as long as all agents have the same class i.e. homogenous preferences (as Table 3 shows). We show here even a stronger result, that is even when agents preferences are heterogeneous, and no matter what the combination of classes of preferences that they have, the skeptical operator sustains Pareto optimality. This establishes the robustness of the skeptical operator when it comes to Pareto optimality.

Theorem 6. *Let $\mathcal{R} = \{\ominus, \ominus^{\mathcal{M}}, |\ominus|, \left|\ominus^{\mathcal{M}}\right|\}$. The skeptical operator is Pareto optimal in its respective set given that individuals have heterogeneous preferences from $\mathcal{R}$.*

Proof sketch. Let $\mathcal{L} = so_{\mathcal{AF}}(P)$, for some P. From Theorem 1, $\mathcal{L}$ is the biggest labeling in $\mathcal{E}^{P}_{so}$. Then, $\forall \mathcal{L}' \in \mathcal{E}^{P}_{so} : \mathcal{L}' \sqsubseteq \mathcal{L} \sqsubseteq \mathcal{L}_i, \forall i \in Ag$. From Lemma 1 and Lemma 3, no matter what class of preferences from $\mathcal{R}$ each agent in *Ag* employs, no agent i would prefer a labeling $\mathcal{L}_X$ (that is different from $\mathcal{L}_i$) over $\mathcal{L}$. ■

Given the above results, we realize that the skeptical operator satisfies Pareto optimality given different classes of preferences, while the credulous and super credulous operators can fail to produce Pareto optimal outcomes for some presumed preferences. Specifically, the two operators fail to do that when agents are assumed to employ distance based preference. Pareto optimality is a very basic property and one should expect that any useful aggregation operator will satisfy it. Thus, these results suggest a strong disadvantage of the two operators. If more committed outcomes are more desirable, there would be a trade-off between choosing operators that produce

more committed outcomes (i.e. credulous and super credulous), and the operator that guarantees Pareto optimal outcomes (i.e. the skeptical operator). On the other hand, in the case of the set based preferences, all operators satisfy Pareto optimality, and so this factor is irrelevant when it comes to choosing an operator.

One might argue that this distinction between set-based preferences and distance based preferences is blurred in reality and has no intuitive meaning. Additionally, since the classes of preferences employed by agents are most probably implicit, it would be impossible to figure out the employed classes of preferences. However, the distinction between set-based preferences and distance based preferences can be meaningful and identifiable in real world applications. Agents that evaluate arguments qualitatively, or to whom arguments are incomparable, can be thought to have set-based preferences. On the other hand, agents that evaluate arguments quantitatively, or to whom the value of arguments are only relevant collectively, can be thought to have distance-based preferences. Whether agents evaluate arguments quantitatively or qualitatively depends to high degree on the context rather than on the private type of the agent. As thus, our results above provide an answer that is relevant within the studied context.

4 Strategy-Proofness

Strategic manipulability is usually an undesirable property in which an agent, upon knowing the preferences of other individuals, has incentive to misrepresent her own true opinion in order to force a collective outcome which is closer to her true opinion. A strategic lie is what an agent can say if and when she has the opportunity to vote strategically.

Definition 24 (Strategic lie)**.** *Let P be a profile and $\mathcal{L}_k \in P$ be the most preferred labeling of an agent with preference $\succeq_k$. Let Op be any aggregation operator. A labeling $\mathcal{L}'_k$ such that $Op(P_{\mathcal{L}_k/\mathcal{L}'_k}) \succ_k Op(P)$ is called a strategic lie, where $P_{\mathcal{L}_k/\mathcal{L}'_k}$ is the profile that results from the profile P after agent k changes her vote from $\mathcal{L}_k$ to $\mathcal{L}'_k$.*

A strategy-proof operator is one where individuals have no incentive to make strategic lies.

Definition 25 (Strategy-proof operator)**.** *An aggregation operator Op is strategy-proof if strategic lies are not possible.*

Despite the fact that, as we shall see, for most classes of preference, the aggregation operators turned out to be vulnerable to strategic manipulation, a novel type of lie emerged: the benevolent lie. Unlike the malicious lie, the benevolent lie has positive effects on some of the other agents and no negative effects on any agent.

Definition 26 (Malicious lie)**.** *Let Op be some aggregation operator and P be a profile of labelings. We say that a strategic lie $\mathcal{L}'_k$ is malicious iff, for some agent $j \neq k$, $Op(P) \succ_j Op(P_{\mathcal{L}_k/\mathcal{L}'_k})$.*

Definition 27 (Benevolent lie)**.** *Let Op be some aggregation operator and P be a profile of labelings. We say that a strategic lie $\mathcal{L}'_k$ is benevolent iff, for any agent i $Op(P_{\mathcal{L}_k/\mathcal{L}'_k}) \succeq_i Op(P)$ and there exists an agent $j \neq k$, $Op(P_{\mathcal{L}_k/\mathcal{L}'_k}) \succ_j Op(P)$.*

4.1 Connections between Classes of Preferences

Consider an operator Op that only produces labelings that are compatible ($\approx$) with each individual's labeling. The following lemma shows that every strategic lie with the operator Op given IUO Hamming distance based preferences is also a strategic lie given Hamming distance based preferences. This lemma is crucial to show that the benevolence property of lies with the skeptical operator carries over from Hamming distance based preferences to IUO Hamming distance based preferences.

Lemma 5. *Let Op be a* compatible operator. *Let $\mathcal{L}_k$ denote the top preference labeling of agent k. Let P be a profile where each agent submits her most preferred labeling, and let $P' = P_{\mathcal{L}_k/\mathcal{L}'_k}$ be a profile that results from P by changing $\mathcal{L}_k$ to $\mathcal{L}'_k$. Let $\mathcal{L}_{Op} = Op_{\mathcal{AF}}(P)$ be the outcome when agent k does not lie. Let $X^k_{|\ominus^{\mathcal{M}}|}$ (resp. $X^k_{|\ominus|}$) be the set of all labelings $\mathcal{L}'_{Op}$ that satisfy the following two properties:*

1. *There exists some labeling $\mathcal{L}'_k$ s.t. $\mathcal{L}'_{Op} = Op_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$ (i.e. $\mathcal{L}'_{Op}$ is a possible outcome given some lie by agent k), and*
2. *$\mathcal{L}'_{Op} \succ_{k,|\ominus^{\mathcal{M}}|} \mathcal{L}_{Op}$ (resp. $\mathcal{L}'_{Op} \succ_{k,|\ominus|} \mathcal{L}_{Op}$).*

Then $X^k_{|\ominus^{\mathcal{M}}|} \subseteq X^k_{|\ominus|}$.

Proof. $\forall \mathcal{L}'_{Op} \in X^k_{|\ominus^{\mathcal{M}}|}$, we have:

1. There exists some labeling $\mathcal{L}'_k$ s.t. $\mathcal{L}'_{Op} = so_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$, and
2. $\mathcal{L}'_{Op} \succ_{k,|\ominus^{\mathcal{M}}|} \mathcal{L}_{Op}$.

We just need to show that $\mathcal{L}'_{Op} \succ_{k,|\ominus|} \mathcal{L}_{Op}$.

Since $\mathcal{L}'_{Op} \succ_{k,|\ominus^{\mathcal{M}}|} \mathcal{L}_{Op}$, then $\mathcal{L}'_{Op} \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_k < \mathcal{L}_{Op} \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_k$. Then:

$$2 \times |\mathcal{L}'_{Op} \ominus^{io} \mathcal{L}_k| + |\mathcal{L}'_{Op} \ominus^{du} \mathcal{L}_k| < 2 \times |\mathcal{L}_{Op} \ominus^{io} \mathcal{L}_k| + |\mathcal{L}_{Op} \ominus^{du} \mathcal{L}_k| \tag{10}$$

Since $|\mathcal{L}_{Op} \ominus^{io} \mathcal{L}_k| = 0$:

$$2 \times |\mathcal{L}'_{Op} \ominus^{io} \mathcal{L}_k| + |\mathcal{L}'_{Op} \ominus^{du} \mathcal{L}_k| < |\mathcal{L}_{Op} \ominus^{du} \mathcal{L}_k| \tag{11}$$

Which implies:

$$|\mathcal{L}'_{Op} \ominus^{io} \mathcal{L}_k| + |\mathcal{L}'_{Op} \ominus^{du} \mathcal{L}_k| < |\mathcal{L}_{Op} \ominus^{du} \mathcal{L}_k| \tag{12}$$

But $|\mathcal{L}'_{Op} \ominus \mathcal{L}_k| = |\mathcal{L}'_{Op} \ominus^{io} \mathcal{L}_k| + |\mathcal{L}'_{Op} \ominus^{du} \mathcal{L}_k|$ and $|\mathcal{L}_{Op} \ominus \mathcal{L}_k| = |\mathcal{L}_{Op} \ominus^{du} \mathcal{L}_k|$. Then:

$$|\mathcal{L}'_{Op} \ominus \mathcal{L}_k| < |\mathcal{L}_{Op} \ominus \mathcal{L}_k| \tag{13}$$

Which means $\mathcal{L}'_{Op} \succ_{|\ominus|} \mathcal{L}_{Op}$. Hence, $\mathcal{L}'_{Op} \in X^k_{|\ominus|}$. ■

We show that the benevolence property carries over from Hamming distance to IUO Hamming distance based preferences.

Theorem 7. *Consider an operator Op that only produces labelings that are compatible ($\approx$) with each individual's labeling. If all strategic lies are benevolent when agents have Hamming distance based preferences then all strategic lies are benevolent when agents have IUO Hamming distance based preferences.*

Proof. Let Op be a *compatible operator*. Let P be a profile, and $\mathcal{L}'_k$ be a strategic lie of agent k. Denote $\mathcal{L}_{Op} = Op_{\mathcal{AF}}(P)$ and $\mathcal{L}'_{Op} = Op_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. From Lemma 1 (1), since the operator Op only produces labelings that are compatible with all individuals' labelings, then for every agent j s.t. $j \neq k$: ($\mathcal{L}_{Op} \succeq_{j,|\ominus|} \mathcal{L}'_{Op}$ iff $\mathcal{L}_{Op} \succeq_{j,|\ominus^{\mathcal{M}}|} \mathcal{L}'_{Op}$) i.e. Hamming distance based preferences and IUO Hamming distance based preferences are equivalent for all agents other than agent k.

Now given Lemma 5, every strategic lie with the operator Op given IUO Hamming distance based preferences is also a strategic lie given Hamming distance based preferences. However, all those lies are benevolent for every agent $j \neq k$ whether she has Hamming distance based preferences or IUO Hamming distance based preferences. Hence, every lie given IUO Hamming distance based preferences is benevolent. ■

Now we turn to studying the strategy-proofness of the three operators: the skeptical, the credulous and the super credulous.

4.2 Hamming Set and Hamming Distance

Following, we show that none of the three operators is strategy-proof given Hamming set (resp. Hamming distance) based preferences.

Observation 2. *The skeptical aggregation operator is not strategy-proof for neither Hamming set nor Hamming distance based preferences. Consider the three labelings in Figure 6. Labeling $\mathcal{L}_1$ of agent 1 when aggregated with $\mathcal{L}_2$ gives labeling $\mathcal{L}_3$, which disagrees with $\mathcal{L}_1$ on all three arguments. But, when agent 1 strategically lies and reports labeling $\mathcal{L}_2$ instead, the result of the aggregation is the same labeling $\mathcal{L}_2$, which differs only on two arguments $\{A,B\}$. The example is valid for both Hamming set and Hamming distance based preferences.*

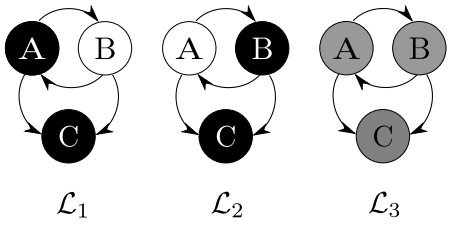

Figure 6: The skeptical operator is not strategy-proof.

Observation 3. *The credulous (resp. super credulous) aggregation operator is not strategy-proof for neither Hamming set nor Hamming distance based preferences. See the example in Figure 7. Labeling $\mathcal{L}_2$ of agent 2 when aggregated with $\mathcal{L}_1$ gives labeling $\mathcal{L}_{CO}$, which disagrees with $\mathcal{L}_2$ on the two arguments. But, when agent 2 strategically lies and reports $\mathcal{L}'_2$ instead, the result of the aggregation is $\mathcal{L}'_{CO}$, which matches the labeling $\mathcal{L}_2$. This lie by agent 2 makes the agent with labeling $\mathcal{L}_1$ worse off. The example is valid for both Hamming set and Hamming distance based preferences.*

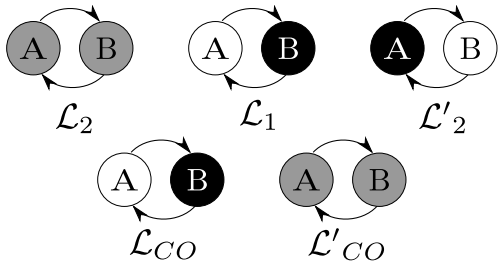

Figure 7: The (super) credulous operator is not strategy-proof.

For the skeptical aggregation operator, however, every strategic lie is benevolent, given Hamming set (resp. Hamming distance) based preferences. Unfortunately, this is not the case for the credulous or the super credulous operators.

Theorem 8. *Consider the skeptical aggregation operator and Hamming set based preferences. For any agent, her strategic lies are benevolent.*

Proof. Let P be a profile, and $\mathcal{L}'_k$ be a strategic lie of agent k. Denote $\mathcal{L}_{SO} = so_{\mathcal{AF}}(P)$ and $\mathcal{L}'_{SO} = so_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. Agent k's preference is $\mathcal{L}'_{SO} \succ_k \mathcal{L}_{SO}$ (i). We will show that for any agent $i \neq k$, we have $\mathcal{L}'_{SO} \succ_i \mathcal{L}_{SO}$. Since the skeptical aggregation operator produces outcomes that are less or equally committed as all the individual labelings, we have that $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$ for all $i \neq k$ (ii). Similarly, we have $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}_k$ (iii). From (i) and (iii), by Lemma 4, we have that $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}'_{SO}$ (iv). From (iv) and (ii) we have $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$ for all $i \neq k$. Finally, we can apply Lemma 3 to obtain $\mathcal{L}'_{SO} \succeq_i \mathcal{L}_{SO}$ for all $i \neq k$ (v). We showed that a lie cannot be malicious, now we show that it is benevolent.

(iii) implies $\texttt{undec}(\mathcal{L}_k) \subseteq \texttt{undec}(\mathcal{L}_{SO})$ (vi). (i) and (vi) imply $\exists A \in \texttt{dec}(\mathcal{L}_k) : A \in \texttt{undec}(\mathcal{L}_{SO}) \wedge A \in \texttt{dec}(\mathcal{L}'_{SO})$ (vii). From (vii), (ii) and (v) $\mathcal{L}'_{SO} \succ_i \mathcal{L}_{SO}$ for $i \neq k$. ■

Theorem 9. *Consider the skeptical aggregation operator and Hamming distance based preferences. For any agent, her strategic lies are benevolent.*

Proof. Let P be a profile, and $\mathcal{L}'_k$ a strategic lie of agent k whose most preferred labeling is $\mathcal{L}_k$. Denote $\mathcal{L}_{SO} = so_{\mathcal{AF}}(P)$ and $\mathcal{L}'_{SO} = so_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. We will show that, if $\mathcal{L}'_{SO}$ is strictly preferred to $\mathcal{L}_{SO}$ by agent k, then it is also strictly preferred by any other agent. Without loss of generality we can take agent $j, j \neq k$,whose most preferred labeling is $\mathcal{L}_j$.

Let us partition the arguments into the following disjoint groups:

- $\mathcal{X} = \texttt{dec}(\mathcal{L}_{SO}) \setminus \texttt{dec}(\mathcal{L}'_{SO})$ (decided arguments that became undecided).

- $\mathcal{Y} = \texttt{dec}(\mathcal{L}'_{SO}) \setminus \texttt{dec}(\mathcal{L}_{SO})$ (undecided arguments that became decided).
- $\mathcal{Z} = \texttt{dec}(\mathcal{L}'_{SO}) \cap \texttt{dec}(\mathcal{L}_{SO})$ (arguments decided in both labelings).
- $\mathcal{V} = \texttt{undec}(\mathcal{L}'_{SO}) \cap \texttt{undec}(\mathcal{L}_{SO})$ (arguments undecided in both labelings).

Labelings $\mathcal{L}_{SO}$ and $\mathcal{L}'_{SO}$ agree on the arguments in $\mathcal{V}$ (which are labeled undec) and $\mathcal{Z}$ (whose arguments are labeled in or out). For the arguments in $\mathcal{Z}$ there are no $\texttt{in} - \texttt{out}$ conflicts between $\mathcal{L}_{SO}$ and $\mathcal{L}'_{SO}$ as the skeptical aggregation operator guarantees social outcomes less or equally committed as $\mathcal{L}_j$. Therefore, only arguments from $\mathcal{X}$ and $\mathcal{Y}$ have an impact on the Hamming distance.

Both labelings $\mathcal{L}_k$ and $\mathcal{L}_j$ agree with $\mathcal{L}_{SO}$ on the arguments in $\mathcal{X}$ because $\mathcal{L}_{SO}$ decides on those arguments and is less or equally committed as both labelings. On the other side, $\mathcal{L}'_{SO}$ remains undecided on the arguments in $\mathcal{X}$ so both labelings $\mathcal{L}_k$ and $\mathcal{L}_j$ disagree with $\mathcal{L}'_{SO}$ on $\mathcal{X}$.

$\mathcal{L}'_{SO}$ is less or equally committed as $\mathcal{L}_j$ so, as above, we obtain that on the arguments in $\mathcal{Y}$, $\mathcal{L}_j$ agrees with $\mathcal{L}'_{SO}$ and disagrees with $\mathcal{L}_{SO}$. On the contrary, $\mathcal{L}'_{SO}$ does not have to be less or equally committed as $\mathcal{L}_k$ and so, for agent k, some of the arguments from $\mathcal{Y}$ increase the distance and some of them decrease. If agent k prefers $\mathcal{L}'_{SO}$ to $\mathcal{L}_{SO}$, then the number of the arguments decreasing the distance must be greater than the number of those increasing by more than $|\mathcal{X}|$. But for agent j all the arguments from $\mathcal{Y}$ are decreasing the distance, as $\mathcal{L}_j$ agrees with $\mathcal{L}'_{SO}$ on the whole $\mathcal{Y}$. So, if agent k gains by switching to labeling $\mathcal{L}'_{SO}$, agent j needs to gain at least the same. ■

Note that the previous two theorems raise an interesting point. Given the Pareto optimality of the skeptical operator for Hamming set/distance based preferences, one would expect that benevolent lies are not possible. Otherwise, it means there exists another labeling that is more preferred by every agent and strictly preferred by at lease one agent. This contradicts the Pareto optimality result found earlier.

However, it is important to remember that the Pareto optimality results found earlier are all with respect to the sets of labelings that are smaller or equal (or compatible in the case of the other operators) to each individuals' labelings. On the other hand, an outcome given a benevolent lie is not compatible with every individual's labeling i.e. while the skeptical operator does produce labelings that are compatible with each individual's true labeling, it does so for the submitted labelings only. Hence, when an agent k lies and submits $\mathcal{L}'_k$ instead of $\mathcal{L}_k$, the outcome $\mathcal{L}'_{SO}$ (which is the outcome when k submits $\mathcal{L}'_k$) is compatible with $\mathcal{L}'_k$ but not necessarily to $\mathcal{L}_k$. As a result, the labeling $\mathcal{L}'_{SO}$ does not belong to the set of labelings that $\mathcal{L}_{SO}$ is compared to when studying Pareto optimality.

This highlights another interesting point that can be implied by the benevolence and Pareto optimality of the skeptical operator. When using the skeptical operator, whenever an agent k considers lying in order to get a closer labeling to $\mathcal{L}_k$, she is faced with an inevitable trade-off between getting a less or equally committed outcome and getting a closer (i.e. more preferred) outcome.

4.3 IUO Hamming Sets and IUO Hamming Distance

In this part, we analyze the strategy-proofness for the three operators given the classes of preferences that assume undec is in the middle between in and out ($dist(\texttt{dec}, \texttt{undec}) < dist(\texttt{in}, \texttt{out})$).

Following, we show the strongest result for this section. The skeptical operator is strategy-proof given the IUO Hamming sets based preferences.

Theorem 10. *The skeptical aggregation operator is strategy-proof when individuals have IUO Hamming sets based preferences.*

Proof. Let P be a profile, $\mathcal{L}_k$ be the top preference of agent k, and $\mathcal{L}'_k \neq \mathcal{L}_k$ be an admissible labeling of $\mathcal{AF}$. Denote $\mathcal{L}_{SO} = so_{\mathcal{AF}}(P)$ and $\mathcal{L}'_{SO} = so_{\mathcal{AF}}(P_{\mathcal{L}_k/\mathcal{L}'_k})$. We will show that $\neg(\mathcal{L}'_{SO} \succ_{k,\ominus^{\mathcal{M}}} \mathcal{L}_{SO})$ (that is, $\mathcal{L}'_k$ is not a strategic lie). Which means, we need to show:

$$\neg((\mathcal{L}'_{SO} \succeq_{k,\ominus^{\mathcal{M}}} \mathcal{L}_{SO}) \wedge \neg(\mathcal{L}_{SO} \succeq_{k,\ominus^{\mathcal{M}}} \mathcal{L}'_{SO})) \tag{14}$$

$$\neg(\mathcal{L}'_{SO} \succeq_{k,\ominus^{\mathcal{M}}} \mathcal{L}_{SO}) \vee (\mathcal{L}_{SO} \succeq_{k,\ominus^{\mathcal{M}}} \mathcal{L}'_{SO}) \tag{15}$$

In other words:

$$\neg((\mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k \subseteq \mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k) \wedge (\mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k \subseteq \mathcal{L}_{SO} \ominus^{du} \mathcal{L}_k)) \vee (\mathcal{L}_{SO} \succeq_{k,\ominus^{\mathcal{M}}} \mathcal{L}'_{SO}) \tag{16}$$

To reformulate, we only need to show that one of the following holds:

1. $\neg(\mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k \subseteq \mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k)$, or
2. $\neg(\mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k \subseteq \mathcal{L}_{SO} \ominus^{du} \mathcal{L}_k)$, or
3.
   (a) $\mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k \subseteq \mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k$, and
   (b) $\mathcal{L}_{SO} \ominus^{du} \mathcal{L}_k \subseteq \mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k$.

First, by definition, $\mathcal{L}_{SO}$ is less or equally committed ($\sqsubseteq$) as $\mathcal{L}_k$. So, $\mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k = \emptyset$, However, this is not the case for $\mathcal{L}'_{SO}$ and $\mathcal{L}_k$. So, $\mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k$ might not be an empty set. Hence, $\mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k \subseteq \mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k$ i.e. (3)(a) is true. Now we show that either (1),(2) or (3)(b) is true.

Suppose (1) and (2) are false and we will show that (3)(b) is then true. This shows that (1), (2), and (3)(b) cannot be all false together.

Since (1) is false and since $\mathcal{L}_{SO} \ominus^{io} \mathcal{L}_k = \emptyset$ then $\mathcal{L}'_{SO} \ominus^{io} \mathcal{L}_k = \emptyset$ (i). Since (2) is false then $\forall a : (a \in \mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k \Rightarrow a \in \mathcal{L}_{SO} \ominus^{du} \mathcal{L}_k)$ (ii). Note that $\forall a : (a \in \mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k \Rightarrow (a \in \texttt{undec}(\mathcal{L}'_{SO}) \wedge a \in \texttt{dec}(\mathcal{L}_k)))$(iii). Otherwise, we would have $a \in \texttt{dec}(\mathcal{L}'_{SO}) \wedge a \in \texttt{undec}(\mathcal{L}_k)$ and from (ii) we would have $a \in \texttt{dec}(\mathcal{L}_{SO}) \wedge a \in \texttt{undec}(\mathcal{L}_k)$ which contradicts $\mathcal{L}_{SO} \sqsubseteq \mathcal{L}_k$.

From (i) and (iii), $\forall a \in \texttt{in}(\mathcal{L}'_{SO}) \Rightarrow a \in \texttt{in}(\mathcal{L}_k)$ (iv) (from (i), $\mathcal{L}_k(a) \neq \texttt{out}$, and from (iii), $\mathcal{L}_k(a) \neq \texttt{undec}$). Similarly, from (i) and (iii), $\forall a \in \texttt{out}(\mathcal{L}'_{SO}) \Rightarrow a \in \texttt{out}(\mathcal{L}_k)$ (v). From (iv) and (v), $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_k$. Since $\forall i \neq k$: $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$, then $\forall i \in Ag$: $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_i$. By Theorem 1, $\mathcal{L}'_{SO} \sqsubseteq \mathcal{L}_{SO}$. Then, $\texttt{undec}(\mathcal{L}_{SO}) \subseteq \texttt{undec}(\mathcal{L}'_{SO})$ (vi).

Now, $\forall a \in \mathcal{L}_{SO} \ominus^{du} \mathcal{L}_k$ then $a \in \texttt{undec}(\mathcal{L}_{SO}) \wedge a \in \texttt{dec}(\mathcal{L}_k)$. From (vi), $a \in \texttt{undec}(\mathcal{L}'_{SO})$. Thus, $a \in \mathcal{L}'_{SO} \ominus^{du} \mathcal{L}_k$. Then, (3)(b) is true. ■

The previous result does not hold for the credulous or the super credulous operators. Further, none of the three operators is strategy-proof when individuals have IUO Hamming distance based preferences. However, as was the case with other classes of preferences, lies with the skeptical operators are always benevolent, unlike those with the credulous or the super credulous operators.

Observation 4. *The skeptical aggregation operator is not strategy-proof when individuals have IUO Hamming distance based preferences. Consider the three labelings in Figure 8. Labeling $\mathcal{L}_1$ of agent 1 when aggregated (using skeptical operator) with $\mathcal{L}_2$ gives labeling $\mathcal{L}_3$, which differs from $\mathcal{L}_1$ on all five arguments with respect to* $\mathtt{dec}-\mathtt{undec}$ *Hamming set. Then, $\mathcal{L}_1 \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_3 = 2\times 0+1\times 5=5$. But, when the agent strategically lies and reports labeling $\mathcal{L}_2$ instead, the result of the aggregation is the same labeling $\mathcal{L}_2$, which differs only on two arguments $\{A,B\}$ with respect to* $\mathtt{in}-\mathtt{out}$ *Hamming set. Then, $\mathcal{L}_1 \left|\ominus^{\mathcal{M}}\right| \mathcal{L}_2 = 2\times 2+1\times 0=4$.*

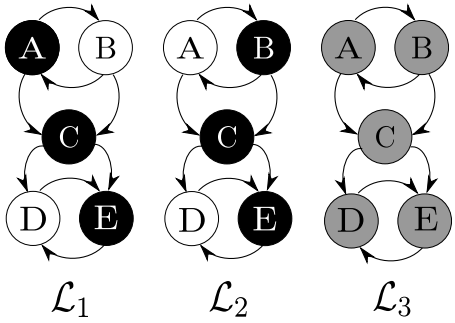

Figure 8: The skeptical operator is not strategy-proof when agents have IUO Hamming distance preferences.

Observation 5. *The credulous (resp. super credulous) aggregation operator is not strategy-proof for neither IUO Hamming sets nor IUO Hamming distance based preferences. The example in Figure 7 can serve as a counterexample for the case where individuals have IUO Hamming sets (or IUO Hamming distance) based preferences. The agent with labeling $\mathcal{L}_2$ can insincerely report $\mathcal{L}_2'$ to obtain her preferred labeling. This makes an agent with labeling $\mathcal{L}_1$ worse off.*

Corollary 5. *Consider the skeptical aggregation operator and IUO Hamming distance based preferences. For any agent, her strategic lies are benevolent.*

Proof. From Theorem 9 and Theorem 7. ■

Table 4 summarizes the strategy-proofness results for the three operators given all the eight classes of preferences.

4.4 Heterogeneous Preferences

Following Subsection 3.4, we do a similar analysis for the case where agents have heterogeneous preferences. Since strategy-proofness is usually considered given other agents' preferences are fixed, it is easy to show the result for the heterogeneous preferences given the homogeneous preferences.

	Skeptical Operator	Credulous Operator	Super Credulous Operator
Hamming set	No (Obs. 2) but benev. (Thm. 8)	No, and not benev. (Obs. 3)	No, and not benev. (Obs. 3)
Hamming dist.	No (Obs. 2) but benev. (Thm. 9)	No, and not benev. (Obs. 3)	No, and not benev. (Obs. 3)
IUO Hamming sets	Yes (Thm. 10)	No, and not benev. (Obs. 5)	No, and not benev. (Obs. 5)
IUO Hamming dist.	No (Obs. 4) but benev. (Cor. 5)	No, and not benev. (Obs. 5)	No, and not benev. (Obs. 5)

Table 4: Strategy-proofness of operators depending on the type of preferences.

Proposition 3. *Let $\mathcal{F}$ be the set of all possible classes of preferences, $\mathcal{R}$ be some set s.t. $\mathcal{R} \subseteq \mathcal{F}$, and Ag be the set of agents. If an operator is strategy-proof given that Ag have homogeneous preferences from $\mathcal{R}$, then it is strategy-proof given that Ag have heterogeneous preferences from $\mathcal{R}$.*

Proof sketch. Since strategy-proofness is considered given that other agents' preferences are fixed, if an agent i has no incentive to lie given some class of preferences, the classes of preferences that are assumed for any agent $k \neq i$ should not affect the incentives of agent i (to lie or otherwise). ■

Given the above results, when using the skeptical operator all strategic lies are benevolent lies, while this is not the case with the credulous and super credulous operators. This introduces another trade-off between choosing the operator that guarantees no malicious lies, and operators that produce more committing outcomes (when more committed outcomes are more desirable).

5 Discussion and Future Work

In order to apply argumentation to multi-agent conflict resolution, it is crucial to take into account not only postulates about logical consistency, but also measures of social optimality and strategic manipulation. Two key criteria are Pareto optimality and strategy-proofness, which are fundamental in any social choice and multi-agent setting. In this study, we have analyzed and compared three aggregation operators, namely the skeptical, the credulous and the super credulous operators with respect to different classes of preferences. Our comparison is based on two fundamental criteria, namely Pareto optimality and strategy-proofness. We showed that the skeptical operator guarantees Pareto optimal outcomes given different classes of preferences, while the credulous and super credulous operators only guarantee Pareto optimal outcomes given a subset of these classes. If more committed outcomes are more desirable, then there is a trade-off between Pareto optimality and the more committed outcomes. As for the strategy-proofness, the three operators are vulnerable to manipulation given most classes of preferences. However, the skeptical operator guarantees benevolent lies. Hence, there is another trade-off in choosing an appropriate operator between avoiding the malicious lies and choosing the more committed outcomes.

All the considered classes of preferences in this work treat arguments independently. Since the label of an argument depends on the labels of the defeating arguments, measuring the distance by treating arguments independently might not give an accurate sense of how far two labelings are from each other. Motivated by this, Booth et al. [13] proposed and defined a new distance method, using the notion of "issue". This distance method captures this idea, while satisfying a set of axiomatic properties which they listed as essential for any distance measure. Given this, it would be sensible to define Issue-wise set and Issue-wise distance measures (and their IUO counterparts). However, agents' preferences defined using (IUO) Issue-wise set turn out to be equivalent to (IUO) Hamming set based preferences. Further, while (IUO) Issue-wise distance based preferences are not equivalent to (IUO) Hamming distance based preferences, all the results found in this paper for (IUO) Hamming distance based preferences turn out to be exactly the same for (IUO) Issue-wise distance based preferences. As such, to avoid a lengthy presentation of similar results, we skipped the presentation of these measures, but the reader can keep in mind that the results found for preferences based on Hamming set and Hamming distance in fact are extensible to broader and potentially more realistic measures. All technical details including definitions and results of Issue-wise related preferences can be found in [5].

The analysis we perform in this study concerns operators that aggregate labelings of an abstract argumentation framework. This problem of aggregating labelings can be compared to preference aggregation (PA) [1, 2, 26, 41], judgment aggregation (JA) [33, 31, 32, 30], and non-binary judgment aggregation [22, 23]. There exist many differences between labelings and preference relations stemming from their corresponding order-theoretic characterizations. Labeling aggregation differ from JA in that arguments (which are the counterparts of propositions) can have three values instead of two traditionally considered in JA. Considering the general framework of Dokow and Holzman [23], our settings can be considered as focusing on special classes of feasible evaluations, which are the conditions imposed by the legal labeling (or other semantics). Additionally, the possible evaluations of each issue (argument, in our case) are to accept (labels as `in`), reject (labels as `out`), or be undecided (labels as `undec`). However, translation of results between labeling aggregation and non-binary JA amounts to encoding argumentation semantics in propositional logic, which is not a trivial task [7, 8].

One might note that the considered operators are insensitive to the number of votes given to each label, which is uncommon in aggregation domains. While most of the common aggregation rules (including the quota rules in JA [20]) are highly dependent on the number of votes received by each alternative, these rules are not always appropriate. One example is in juries, when the legal or the moral responsibility of the outcome is shared by all individuals. Indeed Ronnegard [39] argued that the attribution of moral responsibility to all members of a committee is legitimate when the decision is taken through unanimous voting, while it is not necessarily the case otherwise. Another example is when the outcome of the decision can potentially harm some individuals. It was shown by Bonnefon [11] that people show a preference for more conservative aggregation procedures when the outcome of the decision may involve the infliction of personal harm. The considered three aggregation rules that ensure the compatibility of the outcome with all individuals' votes were proposed to address such scenarios. In a study that experimentally compares the three operators to the argument-wise plurality rule [38, 4], Awad et al. [3] found that while the latter is generally

more favorable, there can be some contextual factors given which this is not the case.

Few studies have considered Pareto optimality and strategy-proofness with argument based aggregation. Rahwan and Larson [36] defined a set of simplistic agents preferences over argumentation outcomes, and studied the Pareto optimality of different argument evaluation rules defined using classical semantics (e.g. *complete*,...etc.) given agents with these simple types of preferences. Unlike Rahwan and Larson, we study the Pareto optimality of labeling aggregation operators that produce a collective evaluation given many different evaluations. Another difference is that we consider more realistic, distance-based preferences. As for strategy-proofness, since the Gibbard-Satterthwaite theorem [27, 40], much research has been done towards analyzing strategic manipulation of preference aggregation (PA) rules [28, 29, 34, 18, 35, 25]. Strategy-proofness of judgment aggregation (JA) operators have been first studied by Dietrich and List [19, 21]. In the former, Dietrich mentioned some *independence* conditions that make the rule strategy-proof. In the latter, Dietrich and List showed equivalence between satisfying strategy-proofness and satisfying both the *independence* and *monotonicity* postulates. The first study of strategy-proofness of labeling aggregation operator has been done by Rahwan and Tohmé [38] in the context of a specific labeling aggregation operator (argument-wise plurality rule). They showed the strategy-proofness of this operator given agents with a particular class of preferences, dubbed focal set preferences. Our work considers different labeling aggregation operators, and we provide the first broad analysis for strategy-proofness of labeling aggregation operators given a wide variety of preferences.

Acknowledgements. *Part of this paper was written during a visit of Edmond Awad to the University of Luxembourg which was generously supported by SINTELNET. The research of Martin Caminada was supported by the Engineering and Physical Sciences Research Council (EPSRC, UK), grant ref. EP/J012084/1 (SAsSY project). The research of Gabriella Pigozzi benefited from the support of the project AMANDE ANR-13-BS02-0004 of the French National Research Agency (ANR). The research of Mikołaj Podlaszewski was supported by the National Research Fund, Luxembourg (LAAMIcomp project).*

References

[1] Kenneth J. Arrow. *Social choice and individual values*. Wiley, New York NY, USA, 1951.

[2] Kenneth J. Arrow, Amartya Sen, and Kotaro Suzumura, editors. *Handbook of Social Choice and Welfare*, volume 1. Elsevier Science Publishers (North-Holland), 2002.

[3] Edmond Awad, Jean-François Bonnefon, Martin Caminada, Thomas W. Malone, and Iyad Rahwan. Experimental assessment of aggregation rules in argumentation-enabled collective intelligence. *ACM Transactions on Internet Technology*, (in press), 2017.

[4] Edmond Awad, Richard Booth, Fernando Tohmé, and Iyad Rahwan. Judgment aggregation in multi-agent argumentation. *Journal of Logic and Computation*, 27(1):227–259, 2017.

[5] Edmond Awad, Martin Caminada, Gabriella Pigozzi, Mikołaj Podlaszewski, and Iyad Rahwan. Pareto optimality and strategy proofness in group argument evaluation (extended version). *arXiv preprint arXiv:1604.00693*, 2016.

[6] Trevor J. M. Bench-Capon and Paul E. Dunne. Argumentation in artificial intelligence. *Artificial Intelligence*, 171(10–15):619–641, 2007.

[7] Philippe Besnard and Sylvie Doutre. Checking the acceptability of a set of arguments. In *NMR*, volume 4, pages 59–64, 2004.

[8] Philippe Besnard, Sylvie Doutre, and Andreas Herzig. Encoding argument graphs in logic. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 345–354. Springer, 2014.

[9] Philippe Besnard and Anthony Hunter. *Elements of Argumentation*. MIT Press, Cambridge MA, USA, 2008.

[10] Gustavo Bodanza, Fernando Tohmé, and Marcelo Auday. Collective argumentation: A survey of aggregation issues around argumentation frameworks. *Argument & Computation*, (Preprint):1–34.

[11] Jean-François Bonnefon. Behavioral evidence for framing effects in the resolution of the doctrinal paradox. *Social choice and welfare*, 34(4):631–641, 2010.

[12] Richard Booth, Edmond Awad, and Iyad Rahwan. Interval methods for judgment aggregation in argumentation. In *Proceedings of the 14th International Conference on Principles of Knowledge Representation and Reasoning (KR 2014)*, pages 594–597, 2014.

[13] Richard Booth, Martin Caminada, Mikołaj Podlaszewski, and Iyad Rahwan. Quantifying disagreement in argument-based reasoning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 493–500. International Foundation for Autonomous Agents and Multiagent Systems, 2012.

[14] Martin Caminada. On the issue of reinstatement in argumentation. In Michael Fisher, Wiebe van der Hoek, Boris Konev, and Alexei Lisitsa, editors, *Proceedings of the 10th European Conference on Logics in Artificial Intelligence (JELIA)*, volume 4160 of *Lecture Notes in Computer Science*, pages 111–123. Springer, 2006.

[15] Martin Caminada and Richard Booth. A dialectical approach for argument-based judgment aggregation. In P. Baroni, Th.F. Gordon, T. Scheffler, and M. Stede, editors, *Computational Models of Argument; Proceedings of COMMA 2016*, pages 179–190. IOS Press, 2016.

[16] Martin Caminada and Gabriella Pigozzi. On judgment aggregation in abstract argumentation. *Autonomous Agents and Multi-Agent Systems*, 22(1):64–102, 2011.

[17] Martin Caminada, Gabriella Pigozzi, and Mikołaj Podlaszewski. Manipulation in group argument evaluation. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume One*, pages 121–126. AAAI Press, 2011.

[18] John R Chamberlin. A mathematical programming approach to assessing the manipulability of social choice functions. *Political Methodology*, 8(4):25–38, 1982.

[19] Franz Dietrich. Judgment aggregation:(im) possibility theorems. *Journal of Economic Theory*, 126(1):286–298, 2006.

[20] Franz Dietrich and Christian List. Judgment aggregation by quota rules: Majority voting generalized. *Journal of Theoretical Politics*, 19(4):391–424, 2007.

[21] Franz Dietrich and Christian List. Strategy-proof judgment aggregation. *Economics and Philosophy*, 23:269–300, 2007.

[22] Elad Dokow and Ron Holzman. Aggregation of binary evaluations with abstentions. *Journal of Economic Theory*, 145(2):544–561, 2010.

[23] Elad Dokow and Ron Holzman. Aggregation of non-binary evaluations. *Advances in Applied Mathematics*, 45(4):487–504, 2010.

[24] Phan M. Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and *n*-person games. *Artificial Intelligence*, 77(2):321–358, 1995.

[25] Pierre Favardin, Dominique Lepelley, and Jérôme Serais. Borda rule, Copeland method and strategic manipulation. *Review of Economic Design*, 7(2):213–228, 2002.

[26] Wulf Gärtner. *A Primer on Social Choice Theory*. Oxford University Press, 2006.

[27] Allan Gibbard. Manipulation of voting schemes. *Econometrica*, 41:587–601, 1973.

[28] Allan Gibbard. Manipulation of schemes that mix voting with chance. *Econometrica*, 45:665–681, 1977.

[29] Allan Gibbard. Straightforwardness of game forms with lotteries as outcomes. *Econometrica*, 46:595–614, 1978.

[30] Davide Grossi and Gabriella Pigozzi. *Judgment Aggregation: A Primer*. Morgan & Claypool, 2014.

[31] Christian List. The theory of judgment aggregation: An introductory review. *Synthese*, 187(1):179–207, 2012.

[32] Christian List and Ben Polak. Introduction to judgment aggregation. *Journal of economic theory*, 145(2):441–466, 2010.

[33] Christian List and Clemens Puppe. Judgment aggregation: a survey. In Paul Anand, Clemens Puppe, and Prasanta Pattanaik, editors, *The handbook of rational and social choice*. Oxford University Press, Oxford, UK, 2009.

[34] Hervé Moulin. On strategy-proofness and single peakedness. *Public Choice*, 35(4):437–455, 1980.

[35] Shmuel Nitzan. The vulnerability of point-voting schemes to preference variation and strategic manipulation. *Public Choice*, 47(2):349–370, 1985.

[36] Iyad Rahwan and Kate Larson. Pareto optimality in abstract argumentation. In Dieter Fox and Carla Gomes, editors, *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI-2008)*, pages 150–155, Menlo Park CA, USA, 2008.

[37] Iyad Rahwan and Guillermo R. Simari, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.

[38] Iyad Rahwan and Fernando Tohmé. Collective Argument Evaluation as Judgement Aggregation. In *9th International Joint Conference on Autonomous Agents & Multi Agent Systems, AAMAS'2010, Toronto, Canada*, 2010.

[39] David Rönnegard. *The Fallacy of Corporate Moral Agency*. Springer, 2015.

[40] Mark Satterthwaite. Strategy-proofness and arrow's conditions: Existence and correspondence theorems for voting procedures and social welfare functions. *Journal of Economic Theory*, 10:187–217, 1975.

[41] Philippe Vincke. Aggregation of preferences: a review. *European Journal of Operational Research*, 9(1):17–22, 1982.